



universität
wien

Diplomarbeit

Titel der Diplomarbeit

Testreihenfolgeeffekte bei Persönlichkeits- und
Leistungsdiagnostik in realen Auswahl-situationen:
Ein Experiment in der Personalauswahl

Verfasser

Alexander Leonard Schünemann

Angestrebter akademischer Grad

Magister der Naturwissenschaften (Mag. rer. nat.)

Wien, 2013

Studienkennzahl: 298

Studienrichtung: Psychologie

Betreuerin: Dr. Lale Khorramdel

Vorwort und Danksagung

All jenen Personen gebührt mein tiefster Dank, die mich bei der Erstellung dieser Arbeit auf vielfältige Weise unterstützt haben. Sie waren es, die es mir ermöglicht haben diese Forschungsarbeit zu planen, durchzuführen und auch zu Ende zu schreiben.

An erster Stelle möchte ich Frau Dr. Lale Khorramdel als Wegbereiterin und Betreuerin für ihre Vorarbeit und das Vertrauen danken, mir die Fortführung ihres Forschungsthemas zu ermöglichen. Sowohl menschlich als auch fachlich war es stets eine große Bereicherung mit ihr zusammenzuarbeiten, von ihr zu lernen und von ihrem Wissen zu profitieren. Als Mentorin und Förderin verdanke ich ihr weitaus mehr, als ich an dieser Stelle ausdrücken kann.

Herrn Mag. Simon Lehner danke ich für die Supervision bei der Durchführung dieser Studie. Ohne seine Unterstützung wäre es nicht möglich gewesen, diese Studie in Kooperation mit der ÖBB durchzuführen. Er hat mir gezeigt, wie sich die psychologische Diagnostik im Unternehmenskontext auf hohem Niveau anwenden lässt. Mein Dank gilt auch den Mitarbeitern der Schuhfried GmbH und Herrn Ing. Ernst Brandstätter für die kompetente Unterstützung.

Gedankt sei darüber hinaus den Kolleginnen, die mir bei der Testleitung zur Seite standen, den Teilnehmern des Experiments und allen anderen Wegbegleitern, die auf ihre ganz spezielle und vielleicht unbemerkte Art und Weise auch zu dieser Arbeit beigetragen haben. Ich danke darüber hinaus Liam und Noel Gallagher für die unzweifelhaft großartigste Musik, die jemals geschrieben wurde.

In sowohl emotionaler, motivationaler und auch grammatikalischer Hinsicht möchte ich meiner Mutter für ihre Unterstützung im Verlauf meiner bisherigen bescheiden Lebenszeit danken. Sie war stets so wertvoll wie wesentlich.

Im Interesse einer besseren Lesbarkeit wird in der vorliegenden Arbeit davon abgesehen männliche und weibliche Sprachformen gemeinsam zu verwenden. Es wird stattdessen meist auf die grammatikalisch maskulinen Formen zurückgegriffen. Selbstverständlich versteht sich, dass - wo zutreffend - Personenbezeichnungen stets beiderlei Geschlecht betreffen.

Abstract

Immer häufiger werden bei Personalauswahlprozessen im Unternehmenskontext umfangreiche computerisierte Testbatterien verwendet. Zur Klärung selektionsdiagnostischer Fragestellungen werden dafür unterschiedliche psychologisch diagnostische Verfahren zur Erfassung eines dem Anforderungsprofil entsprechenden Spektrums an Eigenschaften kombiniert. Aus praktischen oder organisatorischen Gründen werden dabei gelegentlich Testreihenfolgen variiert. Mögliche ergebnisrelevante Konsequenzen dieser Praxis bleiben dabei jedoch oftmals unbeachtet – wohl auch, weil bisher nur wenige Studien zu diesem Thema durchgeführt wurden. Dass Reihenfolgeneffekte auf Itemebene ein evidentes Problem darstellen, ist hinlänglich bekannt und gut dokumentiert (vgl. z.B. Knowles, 1988; Smith, 1992; Hartig & Hinrichs, 2002; Hartig, Hölzl & Moosbrugger, 2007). In Hinblick auf Variationen von Testreihenfolgen wurde bereits gezeigt, dass etwa die Bearbeitung kognitiver Leistungstests einen Einfluss auf die darauffolgende Bearbeitung von Persönlichkeitsfragebögen (Hambros, 2002) und Objektiven Persönlichkeitstests (Khorramdel & Frebort, 2011) haben kann. Wiederholt finden sich auch Arbeiten zu den ergebnisrelevanten Konsequenzen einer Reihenfolgenvariation bei Persönlichkeitsfragebögen und Leistungstests (Eiselt, 1991; Eisenhauer, 2008). Die vorliegende Arbeit geht systematisch und zielgerichtet der Fragestellung nach, ob die von Khorramdel und Frebort (2011) und Khorramdel (2010) gefundenen gegensätzlichen Ergebnisse auf die verwendeten Matrizentests zurückgeführt werden können. Hierzu wurde wie in diesen beiden Vorgängerstudien ein Experiment an 70 Bewerbern für ein Trainee-Programm der Österreichischen Bundesbahnen (ÖBB) mit zwei alternativen Testreihenfolgen, bestehend aus kognitiven Leistungstests und objektiven Persönlichkeitstests, durchgeführt. Sämtliche mit den verwendeten Verfahren erhobenen Kennwerte wurden mit einer multivariaten Varianzanalyse auf Gruppenunterschiede untersucht. Ein Haupteffekt der Testreihenfolge konnte nicht nachgewiesen werden ($p = 0,998$). Anhand dieser Ergebnisse wird eine Systematisierung von Testreihenfolgeneffekten anhand unterschiedlicher Typen psychologisch diagnostischer Verfahren und deren Kombination unter Berücksichtigung relevanter Stichprobeneigenschaften erarbeitet.

Schlüsselwörter: Testreihenfolge, Selektionsdiagnostik, Personalauswahl, Objektive Persönlichkeitstest, Leistungstests

Abstract – English

The increasing utilisation of computerised test batteries in personnel selection processes in organisational settings is becoming more evident. A wide variety of psychological assessment tools for selection purposes are available and used to determine the capability and suitability of the applicants according to a predefined requirement profile. Due to practical or organisational reasons, sometimes test sequences within these computer test batteries are varied. Consequences with relevance to the assessment results remain mostly unaccounted for – as of yet only few studies have addressed this topic. There is strong evidence for order effects on item and subtest level in literature (e.g. Knowles, 1988; Smith, 1992; Hartig & Hinrichs, 2002; Hartig, Hölzl & Moosbrugger, 2007). In respect to test sequence it has been shown for example that working on cognitive ability tests influences the performance on subsequently presented personality questionnaires (Hambros, 2002) and objective personality tests (Khorramdel & Frebort, 2011). The influence of order effects on test scores has also been shown repeatedly by variations of test sequences within test batteries consisting of personality questionnaires and ability tests (Eiselt, 1991; Eisenhauer, 2008). The present study systematically investigates whether the inconsistent findings of Khorramdel and Frebort (2011) and Khorramdel (2010) can be attributed to different matrices tests. The present study continues the works of Khorramdel and Frebort (2011) and Khorramdel (2010) by administering two different test orders within a computer test battery consisting of cognitive ability tests and objective personality test in a similar experimental design. The sample consists of 70 applicants for a management-trainee-programme of the Austrian Federal Railways (ÖBB) that took part in an evaluation of their professional potential. All gathered data was analysed by using a multivariate analysis of variance to test for mean differences between experimental groups. No significant main effect of the test order variation could be found ($p = 0,998$). Based on the present studies results and previous findings a systemisation of test order effects depending on the type of assessment tool and the combination of different tools in respect to relevant sample characteristics is proposed.

Key words: test order, psychological assessment, personnel selection, objective personality tests, ability tests

Inhaltsverzeichnis

<i>Vorwort und Danksagung</i>	III
Abstract	V
Abstract – English	VII
I. Einleitung	1
II. Theorieteil	5
1. Begriffsbestimmungen.....	5
1.1. Psychologische Diagnostik	5
1.2. Typen psychologisch diagnostischer Fragestellungen.....	5
1.3. Bestenauswahl versus Auswahl Geeigneter.....	6
2. Psychologische Diagnostik in der Personalauswahl	7
2.1. Exkurs: Personalauswahlprozesse bei der ÖBB – Ein praktisches Beispiel für das Zustandekommen unterschiedlicher Testreihenfolgen in realen Auswahlprozessen	8
3. Gütekriterien und DIN-Norm 33430.....	10
4. Computertests.....	11
5. Adaptives Testen	13
6. Objektive Persönlichkeitstests.....	16
7. Testmotivation	18
8. Kontexteffekte	21
8.1. Antwortverhalten.....	22
8.2. Kontexteffekte bei Reihenfolgenvariationen	23
8.2.1. Carry-over-Effekte	23
8.2.2. Priming-Effekte	24
8.2.3. Lerneffekte	24
8.2.4. Ermüdungseffekte	25
9. Reihenfolgeneffekte	27

9.1.	Itemreihenfolgeeffekte	28
9.1.1.	Itemreihenfolgeeffekte in Persönlichkeitsfragebögen	28
9.1.2.	Itemreihenfolgeeffekte in kognitiven Leistungstests	32
9.1.3.	Itemreihenfolgeeffekte bei adaptiven Verfahren	32
9.2.	Subtestreihenfolgeeffekte	34
9.3.	Testreihenfolgeeffekte	35
9.3.1.	Testreihenfolgeeffekte bei Persönlichkeitsfragebögen	35
9.3.2.	Testreihenfolgeeffekte bei Persönlichkeitsfragebögen und Leistungstests	36
9.3.3.	Testreihenfolgeeffekte bei Persönlichkeitsfragebögen und OPTs	37
9.3.4.	Testreihenfolgeeffekte bei Leistungstests	38
10.	Testreihenfolgeeffekte in der Personalauswahl.....	41
10.1.	Experiment 1: Khorramdel & Frebort (2011).....	42
10.2.	Experiment 2: Khorramdel (2010).....	45
III.	Empirischer Teil.....	50
1.	Ziele der Untersuchung	51
1.1.	Hypothesen	52
2.	Methode	54
2.1.	Untersuchungsplan und Versuchsdesign	54
2.2.	Erhebungsinstrumente.....	55
2.2.1.	AMT	57
2.2.2.	INSBAT	58
2.2.3.	AHA.....	64
2.2.4.	BACO	66
2.2.5.	ILICA	68
2.3.	Durchführung der Untersuchung.....	69
2.4.	Stichprobe	72
3.	Auswertung	75

3.1.	Multivariate Varianzanalyse	76
3.2.	Anmerkung zu den herangezogenen Variablen und Kennwerten.....	79
4.	Ergebnisse	82
4.1.	Deskriptive Statistik	82
4.2.	Voraussetzungsprüfungen der MANOVA	84
4.2.1.	Linearität.....	85
4.2.2.	Multivariate Normalverteilung	86
4.2.3.	Gleichheit der Fehlervarianzen.....	90
4.2.4.	Homogenität der Varianz-Kovarianz-Matrizen	92
4.3.	Ergebnisse der inferenzstatistischen Analysen	93
5.	Interpretation	97
6.	Diskussion.....	100
7.	Zusammenfassung.....	113
IV.	Anhänge	116
1.	Rückmeldebogen	116
2.	Syntax für z-Transformation in SPSS.....	123
3.	Diskriminanzanalyse	123
4.	Screenshots der verwendeten Subtests des BaCO-D und BACO	125
5.	Tabellenverzeichnis.....	127
6.	Abbildungs- und Formelverzeichnis	127
7.	Literaturverzeichnis	128
8.	Lebenslauf.....	141

I. Einleitung

Eine begrüßenswerte Entwicklung der letzten Jahre stellt die zunehmende Professionalisierung von Personalauswahlprozessen in Unternehmen aller Größenordnungen dar. Wo früher oft aufgrund persönlicher Expertise oder aus dem berüchtigten Bauchgefühl heraus wenig nachvollziehbare Personalentscheidungen getroffen wurden, etablieren sich heute neue Vorgehensweisen und Verfahren. Insbesondere durch die Etablierung der DIN 33430 (DIN Deutsches Institut für Normung e.V., 2002) treten objektive Bewertungskriterien für die Beurteilung von Leistungsfähigkeit und Entwicklungspotential immer mehr in den Vordergrund und gewährleisten dort, wo sie angewendet wird, die Nachvollziehbarkeit der Entscheidungsfindung in Personalauswahlprozessen. Transparenz bei der Beurteilung sichert zusätzlich die Akzeptanz von Berufseignungsentscheidungen.

Teil dieser Entwicklung ist es, zur Identifikation geeigneter Bewerber auf wissenschaftlich fundierte psychologisch diagnostische Verfahren zurückzugreifen. Deren korrekte Anwendung gewährleistet ein hohes Maß an Qualität und somit vor allem Fairness im selektionsdiagnostischen Prozess. Professionelle Verfahren sind als Teil einer multimethodalen Herangehensweise in Personalauswahlprozessen hinsichtlich der Güte der gewonnenen Ergebnisse und der daraus abgeleiteten Entscheidungen einem Vorgehen ohne Verwendung wissenschaftlich fundierter und geprüfter Instrumente weit überlegen (Kubinger, 2009).

Voraussetzung für die Verlässlichkeit der Ergebnisse ist jedoch, dass die mit der Durchführung des Auswahlprozesses betrauten Personen für die Anwendung speziell geschult und qualifiziert sind, denn nur so kann das Potenzial psychologisch diagnostischer Verfahren ausgeschöpft werden. Diese Voraussetzung wird nicht ohne Grund als ein zentraler Bestandteil qualitativ hochwertiger Personalauswahlprozesse in der DIN 33430 gefordert. Das Vorgehen im selektionsdiagnostischen Prozess muss gut überlegt und die Verwendung der Instrumente erprobt sein. Der diagnostisch geschulte Psychologe bringt alle erforderlichen Voraussetzungen mit, die es ihm erlauben, die verfügbaren Verfahren auf ihren sinnvollen Einsatz hin zu bewerten und die weitreichenden Konsequenzen der Verwendung nicht nur zu erfassen, sondern auch in größtmöglichem Umfang zu kontrollieren.

Ein Aspekt, dem bisher im diagnostischen Prozess wenig Aufmerksamkeit zukam, ist die Vorgabereihenfolge von psychologisch diagnostischen Verfahren in einer mehrere Verfahren umfassenden Testbatterie. Einige Studien haben diesen Bereich bereits mit dem Ziel der Qualitätssicherung des diagnostischen Prozesses einer kritischen Analyse unterzogen (Khorramdel, 2010; Khorramdel & Frebort, 2011). Es wurde dabei aufgezeigt, dass die unreflektierte und beliebige Variation von Vorgabereihenfolgen eine mögliche Fehlerquelle im diagnostischen Prozess darstellt. Zu möglichen Itemreihenfolgeeffekten innerhalb eines Tests oder Fragebogens liegen bereits umfassenden Forschungsarbeiten vor; diese werden im Theorieteil dieser Arbeit dargestellt. Wer sich etwa mit der Fragebogenentwicklung beschäftigt, wird früher oder später mit der Frage konfrontiert, ob eine inhaltshomogene Itemblockbildung oder doch eine randomisierte Itemreihenfolge zu wählen ist. Veränderungen der Reihenfolge einzelner Items oder ganzer Itemblöcke haben dabei möglicherweise Auswirkung auf die Testleistung, das Verhalten und Erleben der Testperson in der konkreten Testsituation und auf die Akzeptanz der Verfahren (Moosbrugger & Kelava, 2007). Die Aufdeckung der Wirkzusammenhänge verschiedener Itemreihenfolgen in psychologisch diagnostischen Verfahren (Schwarz & Sudman, 1992) führt insbesondere im Hinblick auf die Anwendung dieser Instrumente im Unternehmenskontext zu einigen bedeutsamen Fragen. Wenn die Variation einzelner Items zu ergebnisrelevanten Effekten führt, welchen Einfluss hat die Veränderung der Vorgabereihenfolge von Tests innerhalb umfangreicher Testbatterien? Wie können ergebnisrelevante Auswirkungen kontrolliert und minimiert werden?

Die Variation der Vorgabereihenfolge von Tests wird in Personalauswahlprozessen bisher jedoch regelmäßig praktiziert, was oft sowohl inhaltliche als auch rein praktische Gründe hat. Exemplarisch wird dafür die Vorgehensweise eines realen Personalauswahlprozesses der Österreichischen Bundesbahnen (ÖBB) in der hier vorliegenden Arbeit dargestellt. Im Verlauf dieser Arbeit wird ausgeführt, aus welchen Gründen Reihenfolgevariationen vorgenommen werden, und so die Notwendigkeit aufgezeigt, sich mit dem Einfluss dieser eingehend auseinanderzusetzen.

Eine weitere deutlich zu beobachtende Entwicklung ist die zunehmende Umsetzung und Verwendung fortgeschrittener Testkonzepte. Moderne psychologisch diagnostische Verfahren halten vermehrt Einzug in die selektionsdiagnostische Praxis im Unternehmenskontext. Die exponentielle Steigerung der Rechenleistung in den

letzten Jahren hat beispielsweise erst eine effiziente Konstruktion und Vorgabe adaptiver Testverfahren möglich gemacht. Dass die Verwendung adaptiver Verfahren möglicherweise einen Einfluss auf die Testmotivation (Asseburg, 2011) und die Leistung von Testpersonen in darauffolgenden Tests hat, lassen beispielsweise die Arbeiten von Khorramdel und Frebort (2011) und Khorramdel (2010) vermuten. Da sich die Ergebnisse der beiden genannten Studien allerdings nicht deckten wurde von den Autoren die Vermutung aufgestellt, dass die Verwendung eines adaptiven Matrizentests in einer der beiden Studien einen Einfluss auf nachfolgende Testergebnisse in Objektiven Persönlichkeitstests hat. Ein Ziel der vorliegenden Studie war es, diesen möglichen Einfluss nochmals zu untersuchen.

Die Verwendung moderner Testkonzepte und computerisierter Verfahren ermöglicht es, die experimentelle Verhaltensbeobachtung als Instrument zur Persönlichkeitsdiagnostik zu verwenden - ein Ansatz, der auch in der vorliegenden Arbeit im Fokus steht. Eine neue Generation computerisierter objektiver Persönlichkeitstests hat das Potenzial die traditionell immer noch in flächendeckender Verwendung befindlichen Persönlichkeitsfragebögen in Auswahl-situationen abzulösen. Da objektive Persönlichkeitstests vergleichsweise noch zu einer eher jungen Generation von Tests zählen, gilt es diese Instrumente ausführlich und unter Einhaltung wissenschaftlicher Standards zu erproben. Der große Vorteil dieser Verfahren liegt in der Vermeidung der evidenten und bisher kaum zu verhindernden Problematik der Verfälschbarkeit klassischer Persönlichkeitsfragebögen. Objektive Persönlichkeitstests befinden sich bereits seit längerem am Markt, leider ist der Anklang dieser Verfahren in der Praxis bisher vergleichsweise gering. Durch die zunehmende Ökonomisierung von Personalauswahlprozessen stellen diese neuen Ansätze und Verfahren allerdings ein äußerst interessantes Forschungsgebiet dar, zu dem auch diese Arbeit einen Beitrag leistet. Im Rahmen dieser Studie wurden auch Objektive Persönlichkeitstests (OPT) einer kritischen Überprüfung unter realen Bedingungen in einem wissenschaftlichen Experiment unterzogen.

Die vorliegende Forschungsarbeit stellt die dritte Studie in einer Reihe von Forschungsarbeiten dar (Khorramdel & Frebort, 2011; Khorramdel, 2010), die sich mit der Variation der Reihenfolge von computerisierten objektiven Persönlichkeitstests und kognitiven Leistungstests im Rahmen realer Auswahl-situation beschäftigen. Das Verhalten realer Bewerber, denen persönlich relevante Konsequenzen aus einer Testung entstehen, unterscheidet sich in erheblichem Maße von freiwilligen

Testteilnehmern (Karner, 2002; Kubinger, 2009). Es ist somit in besonderem Maße relevant, dass auf die Praxis anzuwendende Forschungsergebnisse auch tatsächlich unter praxisnahen - heißt persönliche Konsequenzen zeitigenden - Bedingungen erhoben werden. Diesem Umstand wird in der vorliegenden Studie ebenfalls Rechnung getragen, da die Testung im Rahmen einer realen Auswahl-situation durchgeführt wurde. Die herangezogene Stichprobe liefert dementsprechend hochwertige und realistische Daten und Ergebnisse.

Die vorliegende Arbeit profitiert von der Möglichkeit, die gewonnenen Erkenntnisse aufgrund weitgehend vergleichbarer Versuchsdesigns mit den beiden vorausgegangenen Studien vergleichen zu können. Die Ergebnisse sind dabei lediglich auf reale Personalauswahlprozesse übertragbar. Gerade diese hohe ökologische Validität ist es allerdings, welche die in dieser Arbeit gewonnenen Erkenntnisse für den Anwender im Unternehmenskontext bedeutsam machen.

II. Theorieteil

Zur Auseinandersetzung mit der vorliegenden Arbeit gilt es zunächst Einigkeit über die Terminologie der psychologischen Diagnostik im Bereich der Personalauswahl herzustellen. In diesem einleitenden Abschnitt werden die für das Verständnis relevanten Begrifflichkeiten, welche in der hier vorliegenden Arbeit Verwendung finden, erläutert.

1. Begriffsbestimmungen

1.1. Psychologische Diagnostik

Als Teilgebiet der Psychologie ist die Psychologische Diagnostik nach Kubinger (2009) jene wissenschaftliche Disziplin, die psychologisches Diagnostizieren für die Praxis vorbereitet. Neben der testtheoretisch geleiteten Verfahrensentwicklung beschäftigt sich der Diagnostiker mit allen theoretischen und praktischen Bereichen des diagnostischen Prozesses. Im Verlauf des diagnostischen Prozesses wird aus einer Forschungsfrage eine konkrete Fragestellung abgeleitet. Anschließend wird anhand eines Anforderungsprofils eine adäquate Verfahrensauswahl getroffen, um diese Fragestellung durch die aus der Bearbeitung gewonnenen Ergebnisse zu beantworten.

1.2. Typen psychologisch diagnostischer Fragestellungen

Es werden zwei im Vorgehen zwar ähnliche aber in ihrer Zielsetzung doch gänzlich unterschiedlich ausgerichtete Typen diagnostischer Fragestellungen unterschieden. In einem förderungsorientierten Ansatz (Kornmann, 2003) werden im Interesse der untersuchten Personen im Verlauf des diagnostischen Prozesses zunächst die für die Fragestellung relevanten Fähigkeiten oder Persönlichkeitsmerkmale erfasst. Daran anschließend werden mögliche Fördermaßnahmen aufgezeigt und mit dem Ziel angewendet, die Entwicklung hin zur Ausschöpfung des individuellen Potenzials der Personen zu begünstigen. Die förderungsorientierte Diagnostik ist fest in der pädagogischen und klinischen Psychologie verankert, erlangt aber auch im Rahmen der Personalentwicklung einen immer höheren Stellenwert. Den untersuchten Personen wird im Rahmen dieses Ansatzes die Chance zur berufsbezogenen Weiterbildung oder auch zur Persönlichkeitsentwicklung eröffnet.

Dem förderungsorientierten Ansatz stehen die eignungs- oder selektionsdiagnostischen Fragestellungen gegenüber. Die Methoden entsprechen formal größtenteils denen des förderungsorientierten Ansatzes. Die Selektionsdiagnostik hat jedoch im Verlauf des diagnostischen Prozesses das Ziel, jene Personen zu identifizieren, die sich überhaupt oder aber am besten für eine in einem Anforderungsprofil klar umschriebene Aufgabe, Tätigkeit oder Position eignen (Woike, 2003). Dieser Ansatz ist also in erster Linie an den Interessen des Auftraggebers orientiert und mündet für die Testteilnehmer in einer Auswahl oder Nichtauswahl.

1.3. Bestenauswahl versus Auswahl Geeigneter

Die Personalauswahl ist als selektionsdiagnostischer Prozess ein Teilbereich der Berufseignungsdiagnostik und beschäftigt sich nach traditionellem Verständnis mit der Bestimmung der am besten geeigneten Bewerber für eine begrenzte Anzahl vakanter Positionen (Schaarschmidt, 2003). Im Rahmen von Personalauswahlverfahren unterscheidet man je nach Zielsetzung - meistens abhängig von Art und Anzahl der zu besetzenden Positionen - ob es sich um eine Bestenauswahl oder um die Auswahl überhaupt geeigneter Personen handelt. Diese Unterscheidung ist in der Selektionsdiagnostik wesentlich, da hierdurch festgelegt wird, ob kriteriums- oder normorientiert vorgegangen werden soll. Liegt eine Situation vor, in der die Anzahl der auszuwählenden Personen nicht kontingiert ist und lediglich anhand eines oder mehrerer festgelegter Kriterien die Eignung einer Person festgestellt werden soll, spricht man von der Auswahl Geeigneter. Das Ziel des selektionsdiagnostischen Prozesses ist dann die zuverlässige Identifikation sämtlicher potentiell geeigneter Personen unter den Bewerbern. Eine solche Situation liegt beispielsweise vor, wenn etwa die Zulassung zu einem Studium von der Erfüllung bestimmter kognitiver oder fachlicher Qualifikationen abhängig ist. In der Personalauswahl üblich ist zumeist eine Situation, in der weniger Positionen als Bewerber zur Verfügung stehen. Ziel des diagnostischen Prozesses ist es dann, die maximale Leistungsfähigkeit der Bewerber zu ermitteln und die präzise Identifikation der im Anforderungsprofil festgelegten Persönlichkeitseigenschaften zu gewährleisten. Anhand der mit psychologisch diagnostischen Verfahren erhobenen Kennwerte werden dann im relativen Vergleich der Bewerber jene ausgewählt, welche die beste Leistung gezeigt, die höchste

Übereinstimmung mit dem Anforderungsprofil aufgewiesen und die beste Prognose über zukünftige Performanz erhalten haben (Wottawa & Hossiep, 1997).

In der Praxis ist der Personalauswahlprozess oft eine Verbindung beider Ansätze, sodass erst jenseits meist mehrerer festgelegter Kriterien überhaupt entschieden wird, wer von den Bewerbern im Vergleich am besten abschneidet. Es werden also die Besten der potentiell Geeigneten gesucht. Der dafür notwendige selektionsdiagnostische Prozess erlangt vor dem Hintergrund einer Verknappung verfügbarer Arbeitsplätze, verbunden mit steigenden Anforderungen an hoch spezialisierte Arbeitskräfte, zunehmende Aufmerksamkeit (Wottawa, 2002).

2. Psychologische Diagnostik in der Personalauswahl

Die Berufseignungsdiagnostik bezeichnet nach Schuler (2006) die Methodologie der Entwicklung, Prüfung und Anwendung eignungsbezogener Erfolgsprognosen und stellt dabei eine wichtige Entscheidungshilfe im beruflichen Kontext dar. Sie ist ein wichtiges Anwendungsfeld der psychologischen Diagnostik, die mit ihrem Methoden- und Verfahrensinventar eigentlich einen festen Platz in Personalauswahlprozessen einnehmen sollte. Dennoch werden die meisten Entscheidungen in diesem Bereich noch zu oft von dafür unzureichend ausgebildeten fachfremden Personen gefällt (Wottawa & Hossiep, 1997). Aus psychologischer Sicht laienhaft getroffene Entscheidungen in der Personalauswahl sind verantwortungslos, bedenkt man die sozialen, wirtschaftlichen und persönlichen Konsequenzen einer Fehlentscheidung. Es ist allerdings auch stets auf den Umstand hinzuweisen, dass die prognostische Validität eines Testergebnisses von Laien oftmals überschätzt wird (Kubinger, 2009).

Der Psychologie gelingt es erst in den letzten Jahren, im Anwendungsfeld der Berufseignungsdiagnostik erneut Fuß zu fassen (Schuler, 2006). Obwohl ein tendenziell steigender Einsatz psychologisch diagnostischer Verfahren (Wottawa, 2002; Schuler, 2006) berichtet wird, bemängeln Wottawa und Hossiep (1997) vor allem den im internationalen Vergleich geringen Prozentsatz der Verwendung dieser Verfahren im deutschsprachigen Raum.

Schaarschmidt (2003) fordert deshalb dezidiert, psychologisch diagnostisches Wissen in ganzer Breite durchzusetzen und das Verfahrensinventar in der

Personalauswahl zu etablieren. Geschieht dies nicht, so werden Personalauswahlprozesse fahrlässig von wissenschaftlicher Expertise abgekoppelt.

Die psychologische Diagnostik stellt drei Ansätze zur Berufseignungsdiagnostik zur Verfügung, die jeweils unterschiedliche Informationen liefern und zur Merkmalerfassung spezifische Methoden verwenden (Schuler & Höft, 2006). Im Eigenschaftsansatz wird mithilfe psychologisch diagnostischer Tests die Erfassung stabiler Persönlichkeits- und Leistungsvariablen angestrebt. Das konkrete Verhalten der Bewerber in einer möglichst realistisch simulierten Arbeitsumwelt steht im Fokus des Simulationsansatzes. Zunehmend auch im Bereich der Personalentwicklung relevant ist der biographische Ansatz. Dabei werden biographischen Daten einer Person mittels Fragebögen oder in Interviews erhoben. Eine professionelle Herangehensweise erfordert in der Regel einen multimethodalen Ansatz. Durch die Kombination der diagnostischen Methoden wird die Abdeckung sämtlicher Facetten der im Anforderungsprofil festgeschriebenen Merkmale erreicht. Die Verwendung mehrerer adäquater psychologisch diagnostischer Verfahren erhöht beispielsweise die Reliabilität der Messung und somit die Qualität des diagnostischen Prozesses insgesamt.

2.1. Exkurs: Personalauswahlprozesse bei der ÖBB – Ein praktisches Beispiel für das Zustandekommen unterschiedlicher Testreihenfolgen in realen Auswahlprozessen

Ein konkretes Beispiel aus der Praxis soll diesen Abschnitt abschließen und darstellen, wie Eignungsdiagnostik als multimethodaler Prozess in Unternehmen konkret aufgebaut sein kann. Die hier wiedergegebenen Informationen wurden im Verlauf der Vorbereitungen der vorliegenden Studie in persönlichen Gesprächen mit den Verantwortlichen der Österreichischen Bundesbahnen (ÖBB) gesammelt und beschreiben den Prozess des Auswahlverfahrens, dessen Teil die vorliegende Studie ist.

Ein dreistufiger Ablauf gliedert das Auswahlverfahren, dabei kommen Methoden aus allen drei Ansätzen der Berufseignungsdiagnostik zur Anwendung. Zur Besetzung vakanter Ausbildungsplätze in einem Trainee-Programm werden die Stellen zunächst intern sowie öffentlich ausgeschrieben. Nach einer Vorauswahl werden die betreffenden Personen zu einem halbstandardisierten Interview eingeladen, welches

von zwei Verantwortlichen der Personalabteilung geführt wird. Dieser Teil des Auswahlverfahrens entspricht dem Vorgehen des biographischen Ansatzes. Anschließend werden dem Eigenschaftsansatz entstammende Methoden angewendet. Dies geschieht in Form der Vorgabe mehrerer sich ergänzender psychologisch diagnostischer Verfahren, die sowohl den Leistungs- als auch den Persönlichkeitsbereich abdecken.

Hierzu wird an dieser Stelle im Hinblick auf die Forschungsfrage der vorliegenden Arbeit verdeutlicht, aus welchen Gründen im Unternehmenskontext unterschiedliche Testreihenfolgen verwendet werden. Die ÖBB verfügen über mehrere gut ausgerüstete Computertesträume, die für so gut wie sämtliche diagnostischen Zwecke verwendet werden. Es werden hier Fragen der Personalauswahl sowie der Personalentwicklung behandelt. Darüber hinaus dienen die Testräume dazu, sämtliche Mitarbeiter, die wie etwa Lokführer speziellen gesetzlichen Bestimmungen für ihre Berufsausübung unterliegen, fortlaufend auf ihre Eignung zu testen. Das Verfahrensinventar der ÖBB setzt sich aus einer großen Auswahl psychologisch diagnostischer Verfahren zusammen, die je nach Fragestellung kombiniert werden. Dabei kommen sowohl kognitive Leistungstests als auch Verfahren zur Persönlichkeitsdiagnostik zum Einsatz. Eine Besonderheit besteht darin, dass für einige Verfahren Eingabemedien wie etwa spezielle Probandentastaturen genutzt werden, diese jedoch nicht auf allen Systemen verfügbar sind. So teilen sich die insgesamt 12 vorhandenen Testplätze jeweils zur Hälfte in Systeme zur Leistungsdiagnostik und zur Persönlichkeitsdiagnostik auf. Beide Systeme werden dementsprechend abwechselnd verwendet, wenn eine diagnostische Fragestellung die Verwendung beider Verfahrensarten erfordert. Die Testpersonen müssen also den Testplatz entsprechend dem Verfahrenstyp wechseln. Oft werden mehrere Personen gleichzeitig getestet, sodass es bei hoher Auslastung der Räumlichkeiten je nach Verfügbarkeit zu einer zufälligen Zuteilung zu den Testsystemen kommt. Unter diesen Umständen wird die Testreihenfolge aufgrund praktischer Umstände quasi zufällig variiert. Die Auswirkungen dieser dann verschiedenen und nicht mehr standardisierten Testsituation werden hier und vermutlich in vielen Unternehmen bisher nicht berücksichtigt.

Im Anschluss an die Interviews und die Testungen entscheidet eine weitere Selektionsrunde darüber, wer zum weiterführenden Assessment-Center eingeladen wird. Diese letzte Phase des Auswahlprozesses entstammt methodisch eindeutig dem

Simulationsansatz und komplettiert so einen mehrstufigen multimethodalen Auswahlprozess wie er von Schuler (2006) und auch von Schaarschmidt (2003) gefordert wird.

3. Gütekriterien und DIN-Norm 33430

Die Bemühungen um einheitliche Standards und Qualitätskriterien für Personalauswahlprozesse resultieren in der Ausarbeitung und Etablierung der DIN-Norm 33430 (DIN Deutsches Institut für Normung e.V., 2002). In diesem Regelwerk werden die Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen anwendungsorientiert dargestellt. Es wird festgelegt, welche Ansprüche an Personalauswahlprozesse gestellt werden müssen, um überhaupt eine den Gütekriterien der wissenschaftlichen Praxis entsprechende diagnostische Entscheidung treffen zu können. Die DIN 33430 bezieht dabei sowohl die im diagnostischen Prozess selbst verwendeten Verfahren als auch die beteiligten Personen mit ein. Es werden Regeln und Anleitungen zur Verfahrensauswahl, zur Auswertung und Interpretation von Testergebnissen und zur erforderlichen Qualifikation der beteiligten Personen festgeschrieben. Als Prozessnorm hat sie die Sicherung einer psychologisch diagnostischen Fundierung von Personalentscheidungen zum Ziel (Kersting, 2003a). Die DIN 33430 orientiert sich an den durch das Testkuratorium der Föderation Deutscher Psychologenvereinigungen (Testkuratorium, 1986) festgelegten Gütekriterien für den psychologisch diagnostischen Prozess. Diese fokussieren sich vornehmlich auf die Qualitätssicherung bei der Entwicklung und Anwendung psychologisch diagnostischer Verfahren. Für die Erfassung dieser Kriterien stehen beispielsweise die von Westhoff et. al. (2010) formulierten Checklisten zur Verfügung. Die DIN 33430 versucht, der Beliebigkeit in den Entscheidungsprozesse bei der Personalauswahl professionelle und transparente Standards gegenüberzustellen. Obwohl die Verbindlichkeit der DIN 33430 nicht gesetzlich vorgeschrieben ist (Reimann, 2010), können sich Unternehmen zur Einhaltung verpflichten und helfen so dabei, die Qualität und Seriosität des Diagnostischen Prozesses zu gewährleisten.

So eindeutig und umfangreich die DIN 33430 auch ist, die in der vorliegenden Arbeit behandelten ergebnisrelevanten Effekte einer Testreihenfolgevariation werden in

dieser Norm nicht aufgegriffen. Es ließe sich eventuell aus der Forderung nach einer größtmöglichen Standardisierung der Testsituation und Verfahrensauswahl ableiten, dass von einer Testreihenfolgenvariation abzusehen sei - konkret erwähnt oder begründet wird dieser Aspekt jedoch nicht.

Die psychologische Diagnostik kennt die drei Hauptgütekriterien Objektivität, Reliabilität und Validität, die durch mehrere Nebengütekriterien wie Eichung, Skalierung, Ökonomie und Nützlichkeit sowie Zumutbarkeit, Unverfälschbarkeit und Fairness ergänzt werden (Lienert & Raatz, 1998). Diese Gütekriterien werden vor allem für die Entwicklung und Beurteilung psychologisch diagnostischer Verfahren, aber auch zur Gewährleistung der Qualität des psychologisch diagnostischen Prozesses insgesamt herangezogen. Im Hinblick auf die vorliegende Arbeit ist das Gütekriterium der Fairness in besonderem Maße hervorzuheben. Nach Kubingers (2009) Definition ist das Gütekriterium der Fairness als erfüllt anzusehen, wenn die aus einem psychologisch diagnostischen Verfahren resultierenden Testwerte zu keiner systematischen Diskriminierung bestimmter Gruppen von Testpersonen führen. Obwohl sich diese Definition auf die natürliche Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen bezieht, lässt sie sich leicht nachvollziehbar auch auf situativ bedingte Gruppenzugehörigkeit erweitern. In Bezug auf die in dieser Arbeit durchgeführte Studie würde also eine Benachteiligung einer Bewerbergruppe durch die Zuordnung zu einer, die Ergebnisse der Auswahltestung negativ beeinflussenden, Testreihenfolge eine Verletzung der Fairness des diagnostischen Prozesses bedeuten. Eine systematische Benachteiligung aufgrund test- und prozessimmanenter Bedingungen durch natürliche, aber auch situative Gruppenzugehörigkeit läuft also eindeutig dem Gütekriterium der Fairness zuwider.

4. Computertests

Bei den in der vorliegenden Studie vorgegebenen psychologisch diagnostischen Verfahren handelt es sich ausschließlich um Computertests. Dieser Umstand ist kritisch in die theoretischen Überlegungen und die Betrachtung der Ergebnisse einzubeziehen, da bei einer computerisierten Vorgabe im Vergleich zu anderen Vorgabevarianten bedeutsame Unterschiede zu beachten sind.

Kubinger diskutierte bereits 1993 anhand verschiedener Gütekriterien die testtheoretische Probleme der Computerdiagnostik und verdeutlichte so, dass der Einzug des Computers in die psychologische Diagnostik keinesfalls triviale Konsequenzen hat. Es ist nicht einfach davon auszugehen, dass vormalig als Papier- und Bleistift-Verfahren vorgegebenen Tests problemlos auf den Computer zu übertragen sind. Zumindest eine Äquivalenzprüfung der Normen ist in jedem Fall erforderlich.

Computerisierte Verfahren garantieren ein hohes Maß an Standardisierung bei der Testvorgabe. Die Objektivität der Messung ist leichter zu gewährleisten, da aufgrund der geringen Interaktion von Testleiter und Testperson die Testleiterunabhängigkeit steigt. Die Auswertungs- und Interpretationsobjektivität ist aufgrund automatisierter Auswertungsalgorithmen gewährleistet (Wagner-Menghin, 2003). Eine höhere Testökonomie wird durch die Zeitersparnis des Testleiters erreicht, da dieser sich während der automatisch ablaufenden Testung anderen Aufgaben widmen kann. Zusätzlich wird durch die technische Kontrolle und die hohe Standardisierung der Testsituation die Fairness und Zumutbarkeit des diagnostischen Prozesses erhöht (Kubinger, 2009). Denkt man beispielsweise an die Exaktheit einer Itemvorgabedauer oder an die Messung von Reaktionszeiten, so werden die Vorteile der technologischen gegenüber der menschlichen Kontrolle deutlich. Obwohl zunächst vermutet, erwies sich die Hoffnung auf eine Verminderung von Verfälschungstendenzen bei der Bearbeitung computerisierter Verfahren jedoch als nicht gerechtfertigt (Menghin & Kubinger, 1996). Somit lässt sich annehmen, dass eine Übertragung von Persönlichkeitsfragebögen in eine computerisierte Form in dieser Hinsicht keine Verbesserung darstellt.

Computertests erlauben eine höhere Flexibilität in der Operationalisierung der zu erfassenden psychologischen Konstrukte (Wagner-Menghin, 2003). Dies kommt etwa bei der Entwicklung objektiver Persönlichkeitstests zum Tragen. Erst durch die Verbreitung leistungsfähiger Computer ist die Etablierung moderner Testkonzepte wie etwa des adaptiven Testens überhaupt realisierbar. Auch die Nutzung spezieller Eingabemedien und dynamischer multimedialer Gestaltungsmöglichkeiten sind als Vorzüge computerisierter Verfahren zu nennen.

Hergovich (1994) konnte für einfache Zeichenübungen zeigen, dass die oftmals vorgebrachten Vorbehalte bezüglich der technischen Unerfahrenheit und einer

möglicherweise daraus resultierenden Überforderung älterer Testpersonen unbegründet sind. Zwar fehlen umfassende Arbeiten über die Benachteiligung technologisch unerfahrener Testpersonen, es zeigt sich jedoch, dass, eine ausreichende Instruktions- und Übungsphase vorausgesetzt, keine Belege für eine Benachteiligung von Testpersonen aufgrund mangelnder technologischer Erfahrung bestehen (Kubinger, 2009). Lediglich in speziellen Fällen werden Effekte berichtet. So wird beispielsweise im Manual der *Arbeitshaltungen* (AHA; Kubinger & Ebenhöf, 2007) für den Kennwert Leistungsniveau eine gute Hand-Auge-Koordinationsfähigkeit als zusätzliche mögliche Interpretation einer hohen Ausprägung in diesem Kennwert genannt. Diese Fähigkeit ist oftmals auf einen erfahrenen Umgang mit den Eingabegeräten des Computers zurückzuführen.

Eine zeitgemäße computerisierte Vorgabe geht mit einer höheren Akzeptanz für die Verfahren bei Testpersonen jüngerer und mittleren Alters einher (Schuler, 2006). Insbesondere wenn der Eindruck von Professionalität vermittelt werden soll, wie es in Personalauswahlsituationen der Fall ist, sind computerisierte Verfahren aufgrund der erhöhten Akzeptanz zu bevorzugen.

5. Adaptives Testen

In der vorliegenden Studie wurden psychologisch diagnostische Verfahren eingesetzt, die den Prinzipien der probabilistischen Testtheorie entsprechend konstruiert wurden. Eine solche Verfahrensentwicklung erlaubt es, Items vorzugeben, die an das jeweilige Leistungsvermögen einer Testperson angepasst sind und so eine adaptive Testung ermöglichen. Im Hinblick auf die Diskussion der Ergebnisse der vorliegenden Studie ist es notwendig, die Auswirkungen dieses Testkonzepts zu erörtern.

Zur Erläuterung des adaptiven Testkonzepts ist zu sagen, dass bei der Entwicklung adaptiver Tests zunächst ein Itempool konstruiert wird, der nach Prüfung auf Konformität mit dem Rasch-Modell kalibriert werden muss. So wird einerseits die Eindimensionalität der Messung garantiert und andererseits können die Itemparameter bestimmt werden. Dabei handelt es sich vereinfacht gesagt um Itemschwierigkeiten. Sind diese bekannt, können anhand einer Schätzung des Fähigkeitsparameters der Testperson Items so ausgewählt werden, dass die Lösungswahrscheinlichkeit für

dieses Item bei der Fähigkeit der Testperson etwa 50% beträgt. Items mit einer Lösungswahrscheinlichkeit von 50% besitzen einen maximalen Informationsgehalt. Zu leichte oder zu schwierige Items enthalten weniger Information, da es bereits im Vorhinein abzusehen wäre, ob die Aufgabe gelöst werden kann oder nicht. Eine Vorgabe solcher Items würde die Testdauer unnötig in die Länge ziehen und ein ineffizientes Verhältnis von Itemanzahl zu Informationsgewinn darstellen. Die Schätzung der Fähigkeitsparameter einer Testperson erfolgt während der Testung fortlaufend nach jedem Item und setzt somit aufgrund des hohen Rechenaufwandes zwingend eine computerisierte Testvorgabe voraus. Das Verfahren passt sich im Verlauf einer Testung mit immer größerer Genauigkeit an die Fähigkeit der Testperson an, ist also adaptiv. Adaptives Testen stellt eine Möglichkeit zur Erhöhung der Testökonomie dar, da lediglich solche Items aus einem Itempool vorgegeben werden, die für jede Testperson antwortabhängig maximalen Informationsgehalt besitzen. In mehreren Studien konnte gezeigt werden, dass bei Leistungstests der Messfehler durch adaptives Testen bei geringerer Itemanzahl im Vergleich zur konventionellen Vorgabe reduziert werden kann (Thissen & Mislevy, 2000; Wainer & Eignor, 2000). Nach Wild (1986) wird bei adaptiver Testvorgabe etwa eine Halbierung der Itemanzahl bei gleichbleibender Messgenauigkeit erreicht. Im besten Fall werden jeder Person nur jene Items vorgegeben, die den größten Informationsgehalt besitzen (van der Linden & Glas, 2000).

Den Versuch einer Quantifizierung monetärer Aspekte der Testökonomiesteigerung durch computerisiertes adaptives Testen nehmen Makransky und Glas (2013) vor. Neben weiteren Belegen für den Vorteil einer Testverkürzung bei gleichbleibender Messgenauigkeit beziffern sie die wirtschaftlichen Vorteile bei der Verwendung adaptiver Tests in Auswahlverfahren mit vielen zehntausend Euro pro Jahr.

Obwohl eine Verkürzung der Testdauer eine Entlastung der Testperson bedeutet und somit positiv zu bewerten ist, stellt sich die Situation in der Praxis deutlich komplexer dar. Frey, Hartig und Moosbrugger (2009) untersuchten den Einfluss adaptiver Testvorgabe auf die Motivation der Testpersonen, und die Ergebnisse deuten auf ein großes Problem der adaptiven Testvorgabe hin: Die wahrgenommene Wahrscheinlichkeit für Erfolg ist bei adaptiver Testvorgabe im Vergleich zur konventionellen Vorgabe mit ansteigenden Itemschwierigkeiten bedeutend geringer. Objektiv ist es ein erheblicher Unterschied, ob eine Testperson die Hälfte der schweren Aufgaben oder im Vergleich dazu die Hälfte der leichten Aufgaben eines Itempools

löst. In der subjektiven Wahrnehmung der Testpersonen jedoch, ist dieser Unterschied nicht wahrnehmbar. In beiden Fällen konnte lediglich etwa die Hälfte der vorgegebenen Aufgaben gelöst werden, das Erleben beider Testpersonen ist somit identisch. Dieser Umstand hat bedeutsame Konsequenzen für die Testmotivation.

Auch wenn Kubinger (2009) leistungsmotivationale Beeinträchtigungen bei adaptiver Testvorgabe für unwahrscheinlich hält, spiegelt dies eine seit langer Zeit tradierte theoretische Ansicht wider, die auf Arbeiten aus den Anfängen der Erforschung adaptiver Testvorgaben zurückgeht (Betz, 1975; Betz & Weiss, 1976). Es wird weithin angenommen, dass die Verkürzung der Testdauer und die Vorgabe ausschließlich fordernder Items bei adaptiven Tests sogar zu einer Motivationssteigerung gegenüber konventionellen Tests führen (Frey, 2006). Erst in jüngerer Zeit werden dieser Annahme auch empirische Arbeiten gegenübergestellt.

Frey, Hartig und Moosbrugger (2009) verglichen eine adaptive und eine nicht-adaptive Variante eines Konzentrationsleistungstests und erklärten die berichtete verminderte Motivation zur Testbearbeitung bei adaptiver Vorgabe mit der subjektiven Einschätzung der Erfolgswahrscheinlichkeit. Bei adaptiver Vorgabe war die Einschätzung der Erfolgswahrscheinlichkeit bedeutend geringer als bei konventioneller Vorgabe. Im Einklang mit diesen motivationsmindernden Effekten einer adaptiven Testvorgabe deuten auch die im Laufe der vorliegenden Erhebung gewonnenen Eindrücke darauf hin, dass die adaptive Testvorgabe ein hohes Maß an Belastung für die Testpersonen darstellte. Dies kann vermutlich auch durch die ungewohnte Art der Testvorgabe begründet werden, die sich doch sehr von der Erfahrung im Umgang mit konventionellen Tests unterscheidet.

Bei der adaptiven Testvorgabe werden konzeptuell bedingt mit großer Regelmäßigkeit, und zwar in etwa 50% der Fälle, Items vorgegeben, die die Testperson nicht lösen kann - ungeachtet ihrer Anstrengungen. Tonidandel, Quiñones und Adams (2002) konnten zeigen, dass die durch die lediglich mittlere Lösungswahrscheinlichkeit ausgelöste Frustration verglichen mit der bei konventioneller Vorgabe deutlich höheren Lösungswahrscheinlichkeit einen negativen Einfluss auf die Testmotivation hat.

Ein weiterer Aspekt leistungsmotivationaler Beeinträchtigung ist die zumeist zu Beginn einer adaptiven Testvorgabe oftmals wahrgenommene „Einpendelung“ der Parameterschätzung und der daraus resultierenden Itemauswahl. Ist der

Standardschätzfehler zu Beginn der Testung noch recht hoch, folgen teilweise extrem schwierige auf extrem leichte Aufgaben. Die Testperson bemerkt dann, wie sich in Abhängigkeit des eigenen Antwortverhaltens sehr leichte und unlösbare Aufgaben abwechseln. Es entsteht somit eine indirekte Rückmeldung über die Korrektheit des Antwortverhaltens. Diese Rückmeldung ist erneut aufgrund des adaptiven Testkonzepts in etwa der Hälfte der Fälle negativ. Dieser Wert liegt weit höher als bei konventioneller Vorgabe. Es wurde auch von Teilnehmern dieser Studie berichtet, dass sich dieser Aspekt negativ auf ihre Testmotivation auswirkte.

Beim adaptiven Testen werden die Testpersonen also im Vergleich zu konventioneller Testvorgabe mit abwechselnd zu leichten und zu schwierigen Aufgaben konfrontiert. Dadurch wird eine Verunsicherung der Testperson ausgelöst. Hinzu kommen die erhöhte Schwierigkeit und die verminderte subjektive Einschätzung der Erfolgswahrscheinlichkeit, die mit adaptiven Testungen einhergehen. Es muss jedoch bedacht werden, dass die motivationalen Auswirkungen auch schlichtweg auf geringe Erfahrung mit diesem Testkonzept zurückzuführen sein können. Aufgrund des hohen Aufwands, der bei der Testkonstruktion entsteht, wird die Entwicklung adaptiver Tests noch oft gescheut, wodurch diese Verfahren entsprechend wenig verbreitet sind. Es mangelt den Testpersonen vermutlich oft einfach an Erfahrung mit diesen Verfahren. Obwohl die Vorteile adaptiver Tests - vor allem durch die Voraussetzung der Gültigkeit des Rasch-Modells - immens sind, setzen sich adaptive Tests erst langsam in der Praxis durch (Ortner, 2008).

6. Objektive Persönlichkeitstests

Vor dem Hintergrund der hohen Verfälschbarkeit von Persönlichkeitsfragebögen, insbesondere in Personalauswahlsituationen, besteht in der psychologischen Diagnostik schon lange das Bestreben, Persönlichkeitseigenschaften auf - im strengeren Sinne von Cattell (1958) - objektive Weise zu erfassen. Gemeint ist hierbei nicht nur Objektivität im Hinblick auf die Verrechnungssicherheit oder Auswertungsobjektivität, sondern auf die Objektivität sämtlicher Aspekte einer Erhebung. Ziel hierbei ist es, einen Test so zu konstruieren, dass er nicht vornehmlich die Leistung, sondern das Verhalten einer Testperson bei der Bearbeitung eines Verfahrens erfasst, um daraus Rückschlüsse auf die Persönlichkeit zu ziehen. Wichtig

ist dabei, dass dies geschieht, ohne dass es der Testperson bewusst wird, in welcher Weise ihr Verhalten die Interpretation der Testergebnisse beeinflusst (Cattell, 1957). Erst unter diesen Voraussetzungen kann eine Erhebung als tatsächlich objektiv bezeichnet werden. Dieser Ansatz basiert nicht auf der Selbstbeschreibungen oder Selbstdarstellung der jeweiligen Person und ist somit weniger anfällig für sämtliche Formen bewusster und unbewusster Verfälschungstendenzen wie etwa dem sozial erwünschten Antwortverhalten. Gerade dieses ist in Personalauswahlsituation in großem Maße zu erwarten. Wobei anzumerken ist, dass das eigentliche Problem nicht das Auftreten von Verfälschungstendenzen selbst ist, sondern der Umstand, dass die Teilnehmer diese Tendenzen in unterschiedlichem Umfang zeigen, womit das statistische herauspartialisieren dieser Effekte psychometrisch massiv erschwert wird. Mit der Etablierung des Computers als Vorgabemedium für psychologisch diagnostische Verfahren wurde es jedoch möglich, den Ansatz objektiver Persönlichkeitstestung als experimentelle Verhaltensbeobachtung zu realisieren. Seit einiger Zeit befinden sich Verfahren im Umlauf, die nach diesem Prinzip entwickelt wurden, es wurden auch in der vorliegenden Studie solche objektiven Persönlichkeitstests (OPT) verwendet.

Auch wenn vielfach gezeigt wurde, dass die Verfälschbarkeit dieser Instrumente im Vergleich zu Persönlichkeitsfragebögen, wenn überhaupt, so lediglich in sehr geringem Maße problematisch ist (Hofmann & Kubinger, 2001; Baldinger, 2006; Kubinger, 2009), spricht Kersting (2003b) in Bezug auf OPTs von einem Akzeptanz-Verfälschbarkeits-Dilemma. Gemeint ist, dass eine Verschleierung der Messintention zwar zu einer geringeren Verfälschbarkeit führt, diese jedoch aufgrund der fehlenden Aufklärung der Testpersonen zu einer verminderten Akzeptanz dieser Verfahren führt. Ortner, Proyer und Kubinger (2006) geben hier jedoch zu bedenken, dass die Akzeptanz von OPTs generell höher ist, wenn sie eine Leistungskomponente enthalten. Die Entwicklung solcher Verfahren ist einigen Einschränkungen unterworfen, die sich vor allem in Schwierigkeiten beim Nachweis der Validität dieser Verfahren bemerkbar machen. Oftmals ist die Operationalisierung psychologischer Konstrukte und persönlichkeitspsychologisch relevanter Konzepte komplizierter als bei anderen Verfahrensgattungen, wodurch die Entwicklung dieser Verfahren erschwert wird. In Ermangelung konstruktnaher Verfahren gestaltet sich eine effektive und verlässliche Validierung der Instrumente oftmals schwierig (Kubinger & Litzenberger,

2003). Dies schränkt wiederum die Brauchbarkeit der Verfahren für die praktische Anwendung erheblich ein.

Dennoch handelt es sich bei den OPTs um eine zukunftssträchtige Gattung psychologisch diagnostischer Verfahren, die unter ständiger wissenschaftlicher und technologischer Weiterentwicklung in naher Zukunft einen essentiellen Part im selektionsdiagnostischen Prozess bei der Bewerberauswahl im Unternehmenskontext einnehmen können (Ortner et al., 2007). Die vorliegende Studie stellt auch in dieser Hinsicht einen Versuch dar, bestehende Verfahren auf ihre Praxistauglichkeit hin zu überprüfen.

7. Testmotivation

Ein weiterer Aspekt, der für die vorliegende Arbeit von großer Bedeutung ist, soll nun aufgegriffen werden: Theorie und Praxis legen nahe, dass es bedeutsame Unterschiede zwischen freiwilligen und nicht freiwilligen Testteilnehmern gibt (Karner, 2002; Kubinger, 2009). Die in der vorliegenden Studie herangezogene Stichprobe ist genau wie in den Vorgängerstudien von Khorramdel (2010) und Khorramdel und Frebort (2011) nicht freiwillig zustande gekommen. Nicht freiwillig meint in diesem Zusammenhang, dass die Teilnehmer die Testung nicht ohne Konsequenzen verweigern konnten, aber auch, dass die Teilnahme an der Testung selbst entweder positive oder negative Konsequenzen von hoher persönlicher Relevanz hat. Dass sich diese Stichproben in bedeutsamer Weise von den üblicherweise ausschließlich Freiwillige umfassenden Stichproben anderer Studien (in den meisten Fällen wohl Studierende bestimmter Fachrichtungen) unterscheiden, erscheint bereits intuitiv wahrscheinlich.

Einleitend soll zur Veranschaulichung auf den Umgang der ÖBB mit dieser Problematik eingegangen werden. Die ÖBB greifen bei ihren Testungen in den meisten Fällen nicht auf die mit den verwendeten Verfahren mitgelieferten Normierungsstichproben zurück, da diese für die von ihnen getesteten realen Teilnehmern oftmals inadäquat sind. Die in der Regel an Freiwilligen gewonnenen Normierungsstichproben stehen oftmals in keinem Verhältnis zum tatsächlichen Leistungsspektrum in realen Auswahl-situationen. In vielen Testungen würden bei der Verwendung der Standardnormen Boden- oder Deckeneffekte auftreten. Daher

werden von der ÖBB nach einer ausreichenden Anzahl von internen Testungen anhand der gesammelten Daten eigene Normtabellen erstellt. Diese werden dann anstelle der Standardnormen verwendet, da diese Standardnormen für praktische Belange bei nicht freiwilligen Testteilnehmern nicht aussagekräftig und damit quasi bedeutungslos sind. Dies verdeutlicht explizit die Unterschiede zwischen freiwilligen und nicht freiwilligen Testteilnehmern in der Praxis und die Auswirkung realer Auswahl-situationen auf die Ergebnisse psychologisch diagnostischer Verfahren.

Zur Erklärung dieser Unterschiede finden sich in der Literatur mehrere Ansätze, die sich vor allem mit motivationalen Aspekten einer Testteilnahme beschäftigen. Es ist anzunehmen, dass die Motivation den entscheidenden Unterschied zwischen freiwilligen und nicht freiwilligen Stichproben darstellt. Im Folgenden werden diesbezüglich einige Ansätze dargestellt.

Heckhausen und Heckhausen (2010) beschreiben etwa den Zusammenhang von Leistungsbedürfnis und Leistungsmotivation. Vereinfacht dargestellt, führt ein durch intrinsische oder extrinsische Motive gefördertes Leistungsbedürfnis zu einer gesteigerten Leistungsmotivation. Moderiert wird dieser Prozess durch das Abwägen der Ergebnisse und die Eintrittswahrscheinlichkeit von Konsequenzen einer Handlung. Dabei ist vor allem deren persönliche Relevanz entscheidend. Übertragen auf die Problematik freiwilliger und nicht freiwilliger Testteilnehmer, ist in einer Testsituation immer dann von einer hohen Testmotivation auszugehen, wenn zunächst das Leistungsbedürfnis hoch ist. Dies kann entweder aufgrund von Persönlichkeitseigenschaften der Fall sein, oder durch den Rahmen, in dem die Testung durchgeführt wird, gegeben sein. Die so entstehende hohe Leistungsmotivation wird zusätzlich durch die persönlich relevanten Konsequenzen verstärkt.

Das von Petty und Cacioppo (1986) beschriebene Elaboration Likelihood Modell (ELM) ist als Erklärungsansatz für die Unterschiede zwischen freiwilligen und nicht freiwilligen Testteilnehmern ebenfalls hilfreich. Laut dem ELM steigt die Wahrscheinlichkeit für eine elaborierte Auseinandersetzung mit einem Thema, in diesem Fall einer Testung im Rahmen eines Auswahlverfahrens, mit dem Grad der persönlichen Bedeutsamkeit, der Verfügbarkeit relevanter Informationen und der persönlichen Fähigkeit. Zusätzlich ist zu einer elaborierten Auseinandersetzung eine Situation mit wenig Ablenkungspotential erforderlich. Personen sind nach dem ELM

vor allem dann motiviert, sich mit etwas auseinanderzusetzen, wenn persönlich relevante Konsequenzen zu erwarten sind und sie sich über die Notwendigkeit einer für die Beschäftigung mit einem Thema erforderlichen kognitiven Anstrengung bewusst sind. Das ELM unterscheidet zwischen einer peripheren und einer zentralen Verarbeitungsrouten, und nur im Falle der zentralen Verarbeitung werden die verfügbaren Ressourcen in ausreichendem Maße ausgeschöpft. Die Motivation zur Teilnahme an den Testungen und der damit einhergehenden Bereitschaft, hohe Leistung zu bringen, hängt also nach dem ELM direkt mit der Bedeutsamkeit für die Testperson zusammen. Von diesem Standpunkt aus kann also für die vorliegende Testung aufgrund der hohen persönlichen Relevanz der Testergebnisse ebenfalls von einer hohen Motivation der Teilnehmer ausgegangen werden.

Auch Schwarz, Hippler und Noelle-Neumann (1992) sehen die Fähigkeit und Motivation zur Verarbeitung als die entscheidenden Determinanten für eine gründliche Auseinandersetzung mit einem Thema - im vorliegenden Fall der Beantwortung eines Items - an. Vorausgesetzt wird dabei, dass die kognitiven Ressourcen und die verfügbaren Gedächtniskapazitäten nicht überstiegen werden. Das Auftreten oder Ausbleiben von Reihenfolgeeffekten bei der Bearbeitung psychologisch diagnostischer Verfahren hängt nach diesem Modell maßgeblich von den verfügbaren kognitiven Ressourcen und der Motivation der Testpersonen ab. Hippler und Schwarz (1987) betonen ebenfalls, dass die Motivation der Testperson der entscheidende Einflussfaktor für die Qualität und Verlässlichkeit der Ergebnisse einer Testung oder der Angaben einer Testperson ist.

Die hier angeführten Arbeiten machen deutlich, dass sich freiwillige und nicht freiwillige Testteilnehmer hinsichtlich ihrer Testmotivation bedeutend unterscheiden. Dieser wichtige und maßgebliche Aspekt wird bisher in den meisten Studien zu Testreihenfolgeeffekten wenig bis gar nicht beachtet. In der vorliegenden Studie ist es gelungen, diesen Umstand zu berücksichtigen.

8. Kontexteffekte

Die vorliegende Arbeit ist thematisch der Erforschung von Kontexteffekten zuzuordnen. Viele Aspekte des diagnostischen Prozesses sind von Kontexteffekten betroffen. Ziel des Diagnostikers ist es, mithilfe psychologisch diagnostischer Verfahren die wahre Eigenschaft oder Fähigkeit einer Person mit einem möglichst geringen Messfehleranteil zu erfassen. Eine Messung ist jedoch aufgrund situativer Einflüsse, aber auch ungenügend erfüllter Gütekriterien der verwendeten Verfahren, wie etwa geringe Reliabilität und Validität, stets mit Messfehlern behaftet. Bei der Verfahrensentwicklung ist man bemüht, den durch die Verfahren verursachten Messfehler so gering wie möglich zu halten und abschätzbar zu machen.

Kontexteffekte sind all jene Einflussfaktoren, die zu unkontrollierbaren Messfehlern führen (Khorramdel, 2010). Hierzu zählen etwa Umwelteinflüsse, Eigenschaften der Testpersonen, die verwendeten Verfahren und besonders die aus diesen Faktoren entstehenden Interaktionseffekte. Die Komplexität des Zusammenwirkens verschiedener Kontexteffekte macht es schwierig, ihren Einfluss eindeutig zu bestimmen. Sind Kontexteffekte und die durch sie verursachten Messfehleranteile eindeutig identifiziert, lassen sie sich oft zumindest rechnerisch kontrollieren. Unerforschte Kontexteffekte führen allerdings zu Messfehleranteilen unbekannten Ursprungs und können somit bei der Interpretation einer Messung nicht adäquat berücksichtigt werden.

Für die hier vorliegende Arbeit sind einige Begrifflichkeiten des breiten Forschungsfeldes zu Kontexteffekten relevant. Da man in der Praxis bei der Personalauswahl vornehmlich am Ergebnis einer Testung orientiert ist - schließlich hängen die zu treffenden Entscheidungen in einer Bewerbungssituation maßgeblich von der Performanz der Teilnehmer ab - konzentriert sich die Erforschung von Kontexteffekten in der Personalauswahl vorwiegend auf jene Effekte, die sich messbar im Ergebnis einer Testung niederschlagen. Sollten nachweislich durch Kontexteffekte verursachte Leistungsunterschiede im Auswahlprozess auftreten, die nicht auf die interindividuellen Fähigkeitsdifferenzen der Testpersonen zurückzuführen sind, ist eine auf diesen Ergebnissen getroffene Entscheidung als unfair anzusehen und im Sinne der DIN 33430 unzulässig.

8.1. Antwortverhalten

Um die Entstehung und die Auswirkungen von Kontexteffekten nachvollziehen zu können, ist ein Verständnis für den Ablauf der Informationsverarbeitung im Antwortprozess notwendig. Tourangeau (1984) und Tourangeau und Rasinski (1988) formulieren in ihren größtenteils auf Einstellungstests und Persönlichkeitsfragebögen basierenden Forschungsarbeiten eine Theorie des Antwortverhaltens. Der von ihnen anhand dieser Verfahren formulierte vierstufige Prozess des Antwortverhaltens ist jedoch leicht auf Leistungstests übertragbar. Bei Leistungstests ist das Antwortverhalten durch das tatsächliche Vorhandensein einer für die Lösung notwendigen Information oder durch die Fähigkeit, die Lösung anhand der in der Aufgabe gegebenen Information zu erarbeiten, moderiert. Für die Beantwortung von Einstellungstests oder Persönlichkeitsfragebögen sind eher Ansichten und Meinungen sowie die Fähigkeit zur Introspektion erforderlich. Davon abgesehen stellt sich der Prozess des Antwortverhaltens in Einstellungstests, Persönlichkeitsfragebögen und Leistungstests in hohem Maße vergleichbar dar. *Tabelle 1* gibt einen Überblick über die vier Stufen des Antwortverhaltens nach Tourangeau und Rasinski (1988), ergänzt durch eine Übertragung auf Leistungstests.

Tabelle 1:

Antwortverhalten bei Persönlichkeitsfragebögen und Leistungstests adaptiert nach Tourangeau und Rasinski (1988)

Stufe 1	Rezeption und Interpretation der Frage- oder Aufgabenstellung
Stufe 2	Aktivierung relevanter Einstellungen und Meinungen oder Aktivierung relevanter Fähigkeits- oder Wissensbereiche
Stufe 3	Anwendung der in Stufe 2 aktivierten Inhalte auf die in Stufe 1 erhaltenen Informationen zur Generierung einer Antwort oder Lösung
Stufe 4	Abgleich der generierten Antwort oder Lösung mit den verfügbaren Antwort- oder Lösungsmöglichkeiten

Der Prozess des Antwortverhaltens beginnt mit der Rezeption des Items, meist einer Frage oder Aufgabe. Die Aufgabenstellung wird interpretiert, worauf die Aktivierung der mit dem Item verknüpften Gedächtnisinhalte erfolgt. Dabei handelt es sich etwa um Einstellungen und Meinungen oder Fähigkeiten und Wissensbestände. Die aus dem Verständnis des Items, der Situation und dem Gedächtnis verfügbaren

Informationen werden daraufhin hinsichtlich Relevanz und Nützlichkeit bewertet und letztlich zur Generierung einer vermeintlichen Lösung oder Antwort integriert. Um nun eine endgültige Antwort zu geben, werden die verfügbaren Lösungsmöglichkeiten mit der selbst generierten abgeglichen und letztlich eine Entscheidung getroffen.

Auf jeder Stufe und an den Übergängen dieses Prozesses üben Kontexteffekte Einfluss auf das Antwortverhalten aus (Smith, 1992). Die Vermutung liegt nahe, dass nicht nur bei jedem Item gesondert betrachtet Kontexteffekte auftreten, sondern dass diese auch itemübergreifend wirken. Die vier Stufen des Antwortverhaltens werden für jedes Item erneut mit durch den Einfluss von Kontexteffekten veränderten Vorzeichen durchlaufen.

8.2. Kontexteffekte bei Reihenfolgenvariationen

Im Folgenden werden nach Khorramdel und Frebort (2011) jene Kontexteffekte angeführt, die von Veränderungen der Vorgabereihenfolge betroffen sind. Im Hinblick auf Variationen von Testreihenfolgen werden *carry-over*-, *Priming*-, Lern- und Ermüdungseffekte genannt.

8.2.1. Carry-over-Effekte

Carry-over- oder Übertragungseffekte sind weit gefasste und sehr bedeutsame Kontexteffekte. Sie umfassen so gut wie sämtliche denkbaren Übertragungsvorgänge von Vorangegangenen auf Folgendes. Die grundlegende Aussage ist, dass bei der Bearbeitung eines Verfahrens stets kausale Beeinflussungen vorliegen und die Bearbeitung einzelner Aufgaben nie vollkommen unabhängig betrachtet werden kann. Jedes Item schafft einen bewussten oder unbewussten Rahmen, vor dessen Hintergrund nachfolgende Items bearbeitet werden (Strack, 1992). Vorangegangene Inhalte, deren Bedeutung sowie durch die Bearbeitung ausgelöste Verhaltensweisen werden auf folgende Items übertragen und beeinflussen das Antwortverhalten (Khorramdel & Frebort, 2011). Die Ausgangssituation einer Testperson ist somit je nach Vorgabereihenfolge vor der Bearbeitung unterschiedlicher Items keinesfalls stets dieselbe.

8.2.2. Priming-Effekte

Priming meint in diesem Zusammenhang die Aktivierung von Gedanken, Ideologien oder Überzeugungen, die Einfluss auf den kognitiven oder emotionalen Zustand haben, der wiederum einen veränderten interpretatorischen Rahmen für die Bearbeitung darauf folgender Items schafft (Hogg & Vaughan, 2008). *Priming*-Effekte können unbewusst sein und indirekt wirken, sie können der Testperson aber auch bewusst sein und so direkten Einfluss auf das Antwortverhalten ausüben (Fazio, Powell & Herr, 1983). Durch unterschiedliches Aufgabenmaterial hervorgerufenen *Priming* kann einerseits bestimmte Lösungsstrategien begünstigen, deren Anwendung aber andererseits auch behindern (La Rue & Olejnik, 1980).

Insbesondere bewusstes *Priming* kann zu Ablehnungseffekten führen. Kontroverse Iteminhalte können reaktantes Antwortverhalten hervorrufen. Ablehnung kann auftreten, wenn der Inhalt eines Items bei der Interpretation berücksichtigt oder die Messintention eines Verfahrens durchschaut und abgelehnt wird. Wenn auch vorwiegend bei Fragebögen zu beobachten, werden Ablehnungseffekte wiederholt in Verbindung mit Objektiven Persönlichkeitstests berichtet. Wenn die Verfahren einen nicht mehr zeitgemäßen Eindruck machen oder die Messintention offensichtlich durchschaubar ist, führt eine geringe Akzeptanz des Itemmaterials zu Ablehnungseffekten und beeinflusst so das Antwortverhalten negativ (Proyer & Ortner, 2004).

8.2.3. Lerneffekte

Lerneffekte stellen sowohl bei Persönlichkeitsverfahren als auch bei Leistungstests bedeutsame Kontexteffekte dar. In beiden Fällen erfolgt im Verlauf der Testung eine Gewöhnung an die Aufgabenstellungen, die konkrete Testsituation, die gestellten Anforderungen und das Antwortformat. Lerneffekte stehen verständlicherweise mit Reihenfolge bzw. Positionseffekten in Verbindung, da es sich beim Lernen um einen Prozess handelt, der im Laufe einer Testung über einen längeren Zeitraum hinweg abläuft.

Hartig, Hölzl und Moosbrugger (2007) zeigen, dass die serielle Position eines Items innerhalb eines Fragebogens einen Einfluss auf die Beantwortung hat. Je weiter hinten ein Item in einem Verfahren steht, desto konsistenter werden die Antworten einer

Person. Knowles (1988) führte dies auf die fortschreitende Durchschaubarkeit der Messintention zurück, wohingegen Hamilton und Shuminsky (1990) dies mit der steigenden Aktivierung und Festigung eines speziellen Selbstschemas erklären.

Die Steigerung der Reliabilitäten bei späterer Itemposition in Persönlichkeitsfragebögen wird von Hartig, Hölzl und Moosbrugger (2007) auf Lerneffekte zurückgeführt. Es kommt zu einer Reduzierung des Messfehlers - der Teilnehmer lernt seine Rolle als Befragter und gewöhnt sich an die Testsituation. Die Testperson gewinnt zunehmend Vertrauen zum Testleiter und zeigt eine höhere Bereitschaft zum Informationsaustausch im Gespräch. Eine verringerte Anspannung führt zu erhöhter Bereitschaft, Auskunft zu geben. Besonders für Persönlichkeitsfragebögen sind die Auswirkungen dieses Verhalten relevant und häufig zu beobachten. Testpersonen festigen ihr Antwortverhalten im Verlauf eines Persönlichkeitsfragebogens, um ein konsistentes Bild ihrer selbst zu präsentieren (Smith, 1992). Dies führt mitunter zu einer verzerrten und von der wahren Eigenschaft abweichenden Messung. Besonders problematisch wird dieser Effekt bei der Verwendung von Persönlichkeitsfragebögen in der Personalauswahl, da auf diese Weise etwa sozial erwünschtes Antwortverhalten begünstigt wird.

8.2.4. Ermüdungseffekte

Ermüdungseffekte sind an formale Aspekte des Verfahrens und der Testsituation gebunden. Sie sind von der absoluten und nicht der relativen Position eines Items, Subtests oder Verfahrens innerhalb des Testverlaufs abhängig. Ermüdungseffekte treten immer dann auf, wenn für eine Person eine Belastung, etwa durch die vorgegebenen Tests oder die Testsituation selbst, entsteht und diese über einen längeren Zeitraum aufrechterhalten wird. Mit fortschreitender Dauer der Testung steigt dabei die interindividuell unterschiedliche Wahrscheinlichkeit, dass sich die Person ihrer körperlichen und mentalen Belastungsgrenzen nähert.

Ermüdung kann sowohl physischer als auch psychischer Natur sein. Es wird zwischen objektiven Ermüdungseffekten, die sich messbar in der Testleistung niederschlagen, und subjektiv wahrgenommener Ermüdung (Muscio, 1921) unterschieden. In Persönlichkeitsfragebögen führen Ermüdungseffekte zu einem vermehrten Auftreten unbeantworteter Items gegen Ende der Bearbeitungszeit (Kraut, Wolfson & Rothenberg, 1975). Bei Leistungstests führen Ermüdungseffekte zu

verlängerter Bearbeitungszeit und erhöhter Fehleranzahl mit steigender Testdauer (Földényi, Tagwerker-Neuenschwander, Giovanoli, Schallberger & Steinhausen, 1999). Die subjektive Wahrnehmung von Ermüdung ist hingegen vor allem durch die Motivation der Testpersonen und deren Einstellung zum Test selbst bedingt und muss sich nicht unbedingt im Ergebnis der Testung niederschlagen (Dodge, 1917).

Ackerman und Kanfer (2009) vermuten, dass die subjektive Wahrnehmung von Ermüdung eher auf stabile Persönlichkeitseigenschaften zurückzuführen ist und nicht durch die Verfahren oder akute Anforderungen verursacht wird. Sind subjektive Ermüdungseffekte eher unabhängig von den verwendeten Verfahren und dem Testsetting, so ist zu vermuten, dass auf Ermüdungseffekte zurückgeführte Reihenfolgeeffekte eher stichprobenspezifisch und wenig generalisierbar sind.

Ein hohes subjektives Empfinden von Ermüdung korreliert nach Ackerman und Kanfer (2009) negativ mit Persönlichkeitseigenschaften wie dem Bedürfnis nach Leistung, dem Verlangen nach Wissen und der Bereitschaft zu intellektueller Anstrengung. In dieser Hinsicht charakterisierte Personen gaben an, weniger Ermüdung zu empfinden, und zeigten zudem geringere Ermüdungseffekte in ihrer Leistungsfähigkeit. Ackerman (2000) berichtet darüber hinaus von einer höheren kognitiven Beanspruchung und somit größerer Anfälligkeit für Ermüdungseffekte durch Tests, die nicht auf fachspezifisches Wissen abzielen, sondern die allgemeine kognitive Leistungsfähigkeit erfassen. In Hinsicht auf die hier vorliegende Studie ist von einer hohen Beanspruchung durch die vorgegebenen Tests auszugehen, da die vorgegebene Testbatterie kein fachspezifisches Wissen, sondern basale arbeitsrelevante kognitive Fähigkeiten erfasst.

Myers (1937) stellt die Vermutung auf, dass subjektive Ermüdungseffekte einen geringen Einfluss auf die tatsächliche Testleistung haben und stattdessen eher ein Ausdruck des Wunsches sind, die Testsituation durch Abbruch des Tests zu verlassen. Ermüdungseffekte führen nach Davis (1946) und Ackerman und Kanfer (2009) vor allem bei leistungsfähigen Testpersonen, denen anspruchsvolle Aufgaben gestellt werden, meist nicht zu einer Verschlechterung der Leistung im Verlauf der Testung, sondern im Gegenteil sogar zu einer Leistungssteigerung. Ackerman und Kanfer (2009) berichten wie auch Kinsman und Weiser (1976), dass die Ermüdungseffekte bei hoher Anforderung durch die Testaufgaben zwar mit physischen Beeinträchtigungen wie Anspannung in Nacken und Schultern sowie

Augenschmerzen einhergehen, diese Beeinträchtigungen sich aber nicht auf die Testergebnisse auswirken. Leistungsfähige Personen sind stattdessen bei hoher Beanspruchung sehr wohl dazu in der Lage, ihre kognitiven Ressourcen auf die gestellten Aufgaben zu fokussieren, Ablenkungen zu vermeiden und bisher ungenutzte Reservekapazitäten auszuschöpfen. Dies führt in Summe dazu, dass die Leistungsfähigkeit über einen langen Zeitraum beibehalten, wenn nicht sogar gesteigert werden kann. Das subjektive Gefühl von Ermüdung tritt ein, bevor die Leistungsfähigkeit tatsächlich durch Ermüdung beeinträchtigt wird (Thorndike, 1900). Auf das subjektive Gefühl von Ermüdung wird dann im Sinne von Kahnemann (1973) und nach Schmidtke (1976) mit der Zuteilung zusätzlicher, bisher zurückgehaltener (Notfall-)Ressourcen reagiert. Die Testperson ist so in der Lage, ihre Leistungsfähigkeit aufrechtzuerhalten.

Positive Aspekte einer Ermüdung im Verlaufe zeitintensiver Testungen werden ebenfalls etwa von Smith (1992) berichtet, auch wenn diese ethisch eventuell zweifelhaft zu sehen sind. Eine Gewöhnung an die Testsituation, auch im Sinne des bereits beschriebenen Lerneffekts, kann zu einem Abbau von Anspannung und Unruhe führen. Eine durch Erschöpfung ausgelöste Resignation kann so die Bereitschaft zur Kooperation und Offenheit erhöhen und in einer Steigerung der Bereitschaft zum Informationsaustausch resultieren.

9. Reihenfolgeneffekte

Die für diese Arbeit grundlegende theoretische Überlegung stammt von Khorramdel (2010) und basiert auf der Annahme, dass die im vorherigen Abschnitt beschriebenen Kontexteffekte bei unterschiedlichen Verfahrensgattungen im Zuge einer Variation der Vorgabereihenfolge auftreten können (Khorramdel & Frebort, 2011). Es wird angenommen, dass diese Effekte sowohl bei Veränderungen der Vorgabereihenfolge von Items innerhalb eines Verfahrens als auch zwischen verschiedenen Subtests eines Verfahrens und in weiterer Folge auch bei der Variation der Reihenfolge innerhalb einer mehrere Verfahren umfassenden Testbatterie auftreten können und dabei einen bedeutsamen Einfluss auf die Testergebnisse haben.

Es werden nun unterschiedliche empirische Forschungsarbeiten zu Reihenfolgeneffekten in psychologisch diagnostischen Verfahren vorgestellt. Die

Forschungsergebnisse werden systematisch jeweils auf den Ebenen Item-, Subtest- und Testreihenfolge dargestellt, und zwar im Hinblick auf die unterschiedlichen Verfahrensgattungen, Persönlichkeitsfragebögen, Leistungstests und Objektive Persönlichkeitstests. So wird eine systematische Übersicht über den Forschungsstand gegeben und eine eindeutige Verortung der vorliegenden Arbeit in dieser Systematik ermöglicht

9.1. Itemreihenfolgeneffekte

Die Erforschung von Reihenfolgeneffekten hat einen starken Bezug zur Konstruktion von Persönlichkeitsfragebogen. Wie im Abschnitt II.8. deutlich wurde, sind diese Effekte in jedem Fall bei der Verfahrensentwicklung zu berücksichtigen (Moosbrugger & Kelava, 2007). Bei der Konstruktion von Persönlichkeitsfragebögen herrschen unterschiedliche Konzepte vor, je nach Messintention der Verfahren wird entweder eine inhaltshomogene Itemblockbildung oder eine systematische Variation der Items verschiedener Skalen und Dimensionen bevorzugt (Rost, 2004).

9.1.1. Itemreihenfolgeneffekte in Persönlichkeitsfragebögen

Die Variation der Itemreihenfolge bei Persönlichkeitsfragebögen erscheint nach Anger (1969) durchaus sinnvoll und wird von Krampen (1993) ausdrücklich empfohlen, da durch Löschungseffekte ein unerwünschter Interpretationsrahmen für nachfolgende Items aufgehoben werden kann. Bei Fragebögen wird durch offensichtlich inhaltshomogene Itemblockbildung die Messintention leicht durchschaubar, dadurch eventuell manipulierbar und etwa für sozial erwünschtes Antwortverhalten anfällig.

Dörrer (2009) verdeutlicht die Unterschiede inhaltshomogener Itemblockbildung und randomisierter Itemreihenfolge hinsichtlich der Dimensionalität eines Persönlichkeitsfragebogens, dem *Wiener-Studieneignungs-Persönlichkeitsinventar* (WSP; Khorramdel & Maurer, unpubliziert). Den Teilnehmern dieser Studie wurden zwei Versionen des Verfahrens mit jeweils geblockter oder randomisierter Itemreihenfolge vorgegeben. Obwohl sich in beiden Varianten vereinzelte Itemparameter unterschieden, konnte kein eindeutiger Effekt einer Itemreihenfolgenvariation identifiziert werden. Es zeigte sich jedoch, dass die

Itemblockbildung der randomisierten Vorgabereihenfolge nicht wie angenommen überlegen ist.

Es gestaltet sich schwierig, eine eindeutige Entscheidung für oder wider die inhaltshomogene Itemblockbildung abzugeben. Krampen, Hense und Schneider (1992) empfehlen etwa, von der beliebten Praxis abzusehen, einzelne Subskalen aus umfassenden Persönlichkeitsfragebögen zu extrahieren, da diese Fragmentierung, die eine Veränderung der Itemreihenfolge bedeutet, zu einer Überschätzung der Reliabilität beitragen könne. Gegenteiliges findet sich bei Rost und Hoberg (1997), die in Bezug auf Itemreihenfolgenvariationen in einem Fragebogen zum schulischen Selbstkonzept keinen Einfluss auf die Konstruktvalidität feststellen konnten und somit keine Bedenken gegen eine Skalenextraktion anmelden.

Eisenhauer (2008) variierte bei zwei Persönlichkeitsfragebögen gleichzeitig die Itemreihenfolge. Neben den Originalversionen wurden zwei weitere Versionen mit umgekehrter Itemreihenfolge sowie inhaltshomogener Itemblockbildung verwendet. Das *Big Five Plus One Persönlichkeitsinventar* (B5PO, Holocher-Ertl, Kubinger und Menghin, 2003) sowie der *Anstrengungsvermeidungstest* (AVT, Rollet & Bartram, 1998) wurden gemeinsam vorgegeben und die Ergebnisse auf Effekte der Itemreihenfolge untersucht. Es konnte für keine der Versionen ein Effekt der Itemposition festgestellt werden; die Mittelwerte der Testpersonen unterschieden sich in keinem Kennwert signifikant.

Zu den konkreten Auswirkungen von Itemreihenfolgeneffekten in Persönlichkeitsfragebögen findet sich eine Vielzahl an Forschungsarbeiten, die vor allem anhand teststatistischer Kennwerte wie Reliabilität und Validität den Einfluss von Itempositionseffekten nachvollziehbar machen. Treten Itempositionseffekte in einem Verfahren auf, führt eine Veränderung der Itemreihenfolge dazu, dass für gleiche Items bei unterschiedlicher Vorgabereihenfolge unterschiedliche Kontexte vorliegen. Die Ergebnisse würden hierdurch verfälscht, eine vergleichende Interpretation wäre nicht mehr zulässig.

Mit einem methodisch komplexen Ansatz untersuchten Hartig, Hölzel und Moosbrugger (2007) Reliabilitätstrends im Verlauf von Persönlichkeitsfragebögen. Die hierzu entwickelte Methode erlaubt es, Veränderungen in der Reliabilität auf Änderungen in der Varianz des wahren Wertes einer Testperson oder aber auf Änderungen der Messfehlervarianz zurückzuführen. Die Ergebnisse, gewonnen aus

einem Selbst- und Fremdbeurteilungsfragebogen zur Selbstwirksamkeitserwartung (von Collani & Schyns, 2002), konnten den Einfluss der absoluten Itemposition in Persönlichkeitsfragebögen anhand von Reliabilitätstrends belegen, die sich in höheren Reliabilitäten gegen Ende einer Testung äußern.

Nicht nur die Reliabilität eines Fragebogens verändert sich im Verlauf einer Testung, sondern auch andere Kennwerte sind von der Itemposition betroffen. In einer Studie von Hartig und Hinrichs (2002) sanken die Itemmittelwerte eines Persönlichkeitsfragebogens in Abhängigkeit von der Anzahl der zuvor bearbeiteten Items. Eine größere Trennschärfe der Items war die Folge, da weniger zustimmend und auch konsistenter geantwortet wurde. Die Autoren erklären dies durch das im Laufe der Bearbeitung eines Fragebogens steigende Verständnis für die den Items zugrunde liegenden Konstrukte. Der Grad des Verständnisses ist je nach Fortschritt der Bearbeitung verschieden und somit mit der Position eines Items in einem Fragebogen verbunden. Bei einer Variation der Reihenfolge wäre somit ein ergebnisrelevanter Effekt des Verständnisgrades die Folge. Die Autoren konnten dies anhand von Veränderungen in den Reliabilitäten und Faktorladungen für unterschiedliche Itemblöcke zeigen. Für gleiche Items gelten bei unterschiedlicher Reihenfolge nicht mehr dieselben Rahmenbedingungen. Eine Vergleichbarkeit der Ergebnisse ist somit auch hier nicht mehr möglich.

Knowles (1988) griff diesen Umstand in zwei Studien auf und konnten für einen Fragebogen zur internalen und externalen Kontrollüberzeugung (James, 1957) serielle Positioneffekte zeigen. Es wurden hierfür Parallelversionen dieses Fragebogens mit vollständiger Rotation der Itempositionen erstellt. Die Ergebnisse konnten erneut die mittlerweile als erwartbar zu bezeichnenden Reliabilitätssteigerungen zeigen und konsistentere Antworttendenzen im Testverlauf belegen. Allerdings erwies sich in diesen Studien der Iteminhalt als weitgehend unbedeutend. Es konnte jedoch eine zunehmende Polarisierung des Antwortverhaltens in Abhängigkeit vom jeweiligen Bearbeitungsfortschritt gezeigt werden. Der Autor hebt in einer anschließenden Veröffentlichung (Knowles et al., 1992) wiederum in besonderem Maße hervor, dass das Erhebungsinstrument - in diesem Falle ein Persönlichkeitsfragebogen - einen bedeutsamen Einfluss auf die Messung selbst hatte. Messung und Instrument sind in hohem Maße voneinander abhängig, und Interaktionseffekte beeinflussen die Ergebnisse bedeutsam. Eine Veränderung des Instruments - beispielsweise durch

eine Veränderung der Itemreihenfolge - verändert somit auch die Messung, da diese nicht unabhängig voneinander sind.

Die in diesem Abschnitt dargestellten Studien zu Reihenfolgeeffekte bei Persönlichkeitsfragebögen verdeutlichen, wie bedeutsam eine Veränderung der Itemreihenfolge für das Antwortverhalten und damit das Ergebnis einer Testung ist. Man kann zusammenfassend sagen, dass ein vielfältiges Spektrum an inhaltlichen und formalen Kontexteffekten bei der Bearbeitung von Persönlichkeitsfragebögen zum Tragen kommt. Obwohl viele der inhaltlichen Effekte wegen ihres komplexen Ursprungs schwer erfassbar bleiben, sind formale Reihenfolgeeffekte vor allem anhand gesteigerter Reliabilitäten, höherer Itemtrennschärfen und konsistenteren Antworten im Verlauf einer Testung feststellbar. Wäre die Itemreihenfolge für das Ergebnis irrelevant, sollte man ungeachtet der seriellen Itemposition von gleichbleibenden Kennwerten ausgehen können. Als besonders problematisch stellt sich die Komplexität und Subjektivität des Erlebens einer einzelnen Testperson in Abhängigkeit der Iteminhalte dar. Kaum abschätzbare oder kontrollierbare Auswirkungen von Kontexteffekten auf die Messergebnisse machen Persönlichkeitsfragebögen in hohem Maße anfällig für Veränderungen der Itemreihenfolge und somit ungeeignet für verlässliche und faire Messungen.

Mehrere Studien liefern Belege dafür, dass Itemreihenfolgeeffekte bei beruflich relevanten Items und Fragen zum leistungsbezogenen Fähigkeitsselbstkonzept in geringerem Maße auftreten (Baehr, 1953; Schriesheim, 1981; Schriesheim, Kopelman, & Solomon, 1989; Rost & Hoberg, 1997; Sparfeldt, Schilling, Rost & Thiel, 2006). Denkt man an die vielfältigen Probleme von Reihenfolgenvariationen bei Persönlichkeitsfragebögen, so scheint es Abweichungen bei leistungsbezogenen Verfahren hinsichtlich möglicher Reihenfolgeeffekte auf Itemebene zu geben.

Itemreihenfolgeeffekte sind bei Persönlichkeitsfragebögen anders charakterisiert und strukturiert als bei der Bearbeitung von Leistungstests. Die Aktivierung unter Umständen kontroverser Inhalte durch Persönlichkeitsfragebögen äußert sich vor allem in inhaltlichen Übertragungseffekten. Iteminhalte von Persönlichkeitsfragebögen können dabei potentiell kontroverser sein als diejenigen von Leistungstests. Im Gegensatz zu Fähigkeit, ist Persönlichkeit in weitaus höherem Maße der subjektiven Wahrnehmung und der emotionalen Introspektion unterworfen.

9.1.2. Itemreihenfolgeneffekte in kognitiven Leistungstests

Bei Leistungstests ist eine Durchmischung der Itemreihenfolge in den seltensten Fällen sinnvoll umsetzbar, so spricht schon die übliche Vorgabe von Übungsaufgaben in der Instruktionsphase eindeutig gegen die Absicht, eine Löschung des Interpretationsrahmens, wie dies etwa bei Persönlichkeitsfragebögen erwünscht sein kann, erreichen zu wollen. Auch eine Verschleierung der Messintention und die Undurchschaubarkeit des Messprinzips sind bei reinen Leistungstests in der Regel kaum wünschenswert oder umsetzbar. Das Problem der Verfälschbarkeit in Personalauswahlsituationen ist bei Leistungstests ebenfalls nicht in hohem Maße und kaum in Richtung eines „faking-good“ evident (Kubinger, 2009).

Nach Rost (2004) empfiehlt sich bei klassischer Leistungstestentwicklung eine Itemreihenfolge in ansteigender Schwierigkeit, um den kognitiven Denkprozess nicht negativ zu beeinflussen. Dies erscheint auch intuitiv richtig und hat darüber hinaus den Vorteil, dass die Testung beendet werden kann, sobald durch eine weiterführende Bearbeitung aufgrund einer Überforderung der Testperson von keinem zusätzlichen Informationsgewinn mehr ausgegangen werden kann.

Perlini, Lind und Zumbo (1998) fanden in einer empirischen Studie, dass Variationen der Itemreihenfolge bei Leistungstests wenig bis keinen Einfluss auf die Testergebnisse haben. Reihenfolgeneffekte bei Leistungstests sind demzufolge eher auf Aspekte wie Testmotivation oder Ermüdung aufgrund kognitiver Anstrengung der Testperson zurückzuführen.

9.1.3. Itemreihenfolgeneffekte bei adaptiven Verfahren

Die Itemreihenfolge wird bei adaptiven Verfahren naturgemäß variiert und weist etwa bei adaptiven Leistungstests einen nicht kontinuierlichen Verlauf des Schwierigkeitsgrades auf. Die Vorgabereihenfolge ist bei adaptiven Verfahren in der Regel von Testperson zu Testperson unterschiedlich.

Itemreihenfolgeneffekte wurden von Ortner (2008) bei adaptiver Vorgabe eines computerisierten Persönlichkeitsfragebogens untersucht. Durch die Manipulation der

Einstiegsfrage eines adaptiven Verfahrens wurden Itemreihenfolgeeffekte hervorgerufen. Unterschiedlich ausgeprägte Einstiegsitems führten zu signifikant höheren respektive niedrigen Ausprägungen in den jeweiligen Skalen.

Leary und Dorans (1985) führen in ihrem Übersichtsartikel mehrere Studien an, die Reihenfolgeeffekte bei adaptiver Testung berichten. Eine der Grundannahmen des Rasch-Modells, welches oft für die Kalibrierung eines Itempools zur Anwendung des adaptiven Testens herangezogen wird, ist die lokale stochastische Unabhängigkeit. Diese Annahme besagt, dass die Beantwortung eines Items unabhängig davon ist, welche Items eine Testperson bereits bearbeitet hat oder noch lösen wird. Es ist demnach unerheblich, an welcher Stelle in einem Test ein Item vorgegeben wird. Die von Leary und Dorans (1985) angeführten Studien und auch die Ergebnisse von Ortner (2004, 2008) deuten auf Probleme bei dieser Annahme hin. Ist die lokale stochastische Unabhängigkeit aufgrund von Itemreihenfolgeeffekten nicht gewährleistet, wird die Parameterschätzung eventuell verzerrt.

Ortner (2004) verglich unterschiedliche Itemreihenfolgen anhand eines dem Rasch-Modell konformen Itempool des Persönlichkeitsfragebogens *Eysenk-Personality-Profiler* (EPP-D, Eysenck, Wilson & Jackson, 1998). Obwohl keine adaptive Version des Fragebogens vorgegeben wurde und auch keine Mittelwertsunterschiede gefunden wurden, zeigten sich trotz Gültigkeit des Rasch-Modells und der damit verbundenen Annahme lokaler stochastischer Unabhängigkeit, bei unterschiedlicher Itemreihenfolge derselben Items dennoch abweichende Itemparameterschätzungen. Es ist somit wahrscheinlich, dass bei einer adaptiven Vorgabe dieses Verfahrens, bei der unterschiedliche Itemreihenfolgen unvermeidbar sind, Itemreihenfolgeeffekte auftreten würden.

Itemreihenfolgeeffekte bei adaptiven Tests treten vorwiegend bei Tests auf, die mit Zeitlimit vorgegeben werden. Handelt es sich um Tests ohne Begrenzung der Arbeitszeit, sind die Parameterschätzungen verlässlich (Leary & Dorans, 1985). Die in der vorliegenden Arbeit vorgegebenen adaptiven Tests enthalten zum Teil Subtests, bei denen die Zeitkomponente einen limitierenden Faktor darstellt.

9.2. Subtestreihenfolgeeffekte

Forschungsarbeiten zu Subtestreihenfolgen sind eher spärlich, es kann hier nur von einigen wenigen Arbeiten berichtet werden. Da unterschiedliche Subtests bei Persönlichkeitsfragebögen kaum vorkommen und eher unter der bereits besprochenen Problematik inhaltshomogener Itemblockbildung laufen, konnten keine Arbeiten hierzu gefunden werden. Subtestreihenfolgen bei OPTs scheinen bisher ebenfalls kaum oder nur unzureichend erforscht worden zu sein. Einzig zu Subtestreihenfolgen bei Leistungstests sind empirische Arbeiten verfügbar.

Schneider und Stern (2010) untersuchten die Bedeutung konzeptuellen und prozeduralen Wissens für das Verständnis von Rechenoperationen bei Kindern. Für diese Studie wurde eine Testbatterie, bestehend aus insgesamt acht Subtests, die jeweils konzeptuelles oder prozedurales Wissen bei der Bearbeitung von Rechenaufgaben erfassen sollen, verwendet. Das Aufgabenmaterial war dabei teilweise aus bestehenden Verfahren oder anderen Studien übernommen oder selbst entwickelt worden. Die Subtests wurden in randomisierter Reihenfolge vorgegeben. Reihenfolgeeffekte traten vereinzelt und vorrangig bei Subtests auf, die ähnliche Konstrukte erfassten. Die Autoren machten letztlich die geringe Validität der verwendeten Subtests für diese Reihenfolgeeffekte verantwortlich.

Im Rahmen einer Normierungsstudie verwendeten Földényi, Tagwerker-Neuenschwander, Giovanoli, Schallberger und Steinhausen (1999) zwei unterschiedliche Formen der *Testbatterie zur Aufmerksamkeitsprüfung* (TAP; Zimmermann & Fimm, 1993). Die Subtests wurden dabei Blockweise variiert. Die resultierenden Reihenfolgen waren wie folgt: Gruppe 1: ABCD – Gruppe 2: CDAB. In fast allen Subtests traten Reihenfolgeeffekte auf. Einige Testpersonen zeigten im Verlauf der Testung abnehmende Reaktionszeiten und begingen zudem mehr Fehler. Andere jedoch steigerten entgegen den Erwartungen der Autoren ihre Leistung bei einigen Aufmerksamkeitsaufgaben, wenn diese gegen Ende der Testung vorgegeben wurden. Die Autoren vermuten Ermüdungseffekte als Ursache für die Leistungsminderung und gaben Lerneffekte als mögliche Gründe für die Leistungssteigerung einiger Testpersonen an. Es ist jedoch zu beachten, dass die herangezogene Stichprobe ausschließlich Kinder im Alter von sechs bis zehn Jahren umfasste. Aufgrund der geringeren Belastbarkeit von Kindern, schränkt dies die Generalisierbarkeit der gefundenen Effekte ein - die Ergebnisse sind für die

vorliegende Arbeit jedoch zumindest dahingehend relevant, dass auch hier dieselbe Form einer blockweisen Variation der Vorgabereihenfolge verwendet wurde. Vor allem Subtests, die Gedächtnisleistungen erfassten, waren betroffen, und es wurde vorwiegend kognitiv forderndes Itemmaterial verwendet. Auch wenn die gefundenen Reihenfolgeeffekte wohl lediglich für die herangezogene Stichprobe charakteristisch sind, ist diese Studie ein Beleg für Reihenfolgeeffekte auf Subtestebene. Gerade weil die Ergebnisse inkonsistent waren, geben die Autoren die abschließende Empfehlung, in der Praxis nach Möglichkeit von einer Variation der Subtestreihenfolge abzusehen.

9.3. Testreihenfolgeeffekte

Khorramdel und Frebort (2011) nehmen an, dass jene Reihenfolgeeffekte, die im Abschnitt II.8. beschrieben wurden, von Item- und Subtestebene auf die Ebene der Testreihenfolgen übertragbar sind. Der folgende Abschnitt umfasst empirische Studien, die zum Thema Testreihenfolgeeffekte durchgeführt wurden. Die hier aufgeführten Studien variieren überwiegend lediglich eine, aber teilweise auch verschiedene Verfahrensgattungen miteinander. Diese Studien geben Anhaltspunkte dafür, welche Effekte bei Variation von Testreihenfolgeeffekten bisher gefunden worden sind und auch in der vorliegenden Studie möglicherweise zu erwarten gewesen wären.

9.3.1. Testreihenfolgeeffekte bei Persönlichkeitsfragebögen

Sowohl Item- als auch Testreihenfolgen bei Persönlichkeitsfragebögen wurden von Eisenhauer (2008) variiert. Im Abschnitt II.9.1.1. wurden bereits die nicht signifikanten Ergebnisse der Reihenfolgenvariation auf Itemebene berichtet. Das B5PO Persönlichkeitsinventar sowie der AVT wurden in mehreren Versionen und in unterschiedlicher Reihenfolge vorgegeben. Auch die Variation der Testreihenfolge führte in lediglich einem Kennwert zu signifikant abweichenden Ergebnissen. Obwohl dieser Kennwert keine Bedeutung für den nicht signifikanten Haupteffekt der Testreihenfolge hatte, wird von der Autorin darauf hingewiesen, dass möglicherweise ein inhaltlicher Übertragungseffekt für diesen Effekt verantwortlich ist. Es deutet sich hier die Anfälligkeit von Persönlichkeitsfragebögen für inhaltliche Kontexteffekte an.

Konsistenzeffekte (Strack, 1994) auf Itemebene könnten sich nach Meinung von Eisenhauer (2008) auf Testreihenfolgenebene ausgewirkt haben. Die Bearbeitung des AVTs habe das Antwortverhalten beim B5PO durch Lerneffekte beeinflusst, da Lerneffekte, wie bereits im Abschnitt II.8. berichtet, bei Persönlichkeitsfragebögen zu konsistenterem Antwortverhalten führen. Ein ähnlicher Effekt wurde auch von Dalbert, Strack, Boesken und Schwenkmezger (1991) bei der Bearbeitung von Bewältigungsfragebögen in klinischen Settings beobachtet. Die Bearbeitung von Bewältigungsfragebögen löste bei darauf folgenden Befindlichkeitsfragebögen signifikant abweichende Angaben aus. Einschränkend ist zu beachten, dass klinische Stichproben nur eine limitierte Generalisierung der Ergebnisse erlauben.

9.3.2. Testreihenfolgeneffekte bei Persönlichkeitsfragebögen und Leistungstests

Hambros (2002) führte eine Studie zur Zumutbarkeit zweier Persönlichkeitsfragebögen mit Fokus auf deren unterschiedliche polytome und dichotome Antwortformate durch. Die Vorgabereihenfolge der Persönlichkeitsfragebögen wurde gemeinsam mit zwei kognitiven Leistungstests variiert. Die vorgegebenen Leistungstests - der *Zahlenverbindungstest* (ZVT; Oswald & Rothe, 1987) und das *Frankfurter Aufmerksamkeits Inventar* (FAIR, Moosbrugger & Oehlschlägel, 1996) - wurden von den Autoren dezidiert deshalb verwendet, weil diese Verfahren bei der Bearbeitung lediglich geringer kognitiver Anstrengung bedürfen.

Hambros (2002) nahm an, dass das dichotome Antwortformat des Persönlichkeitsfragebogens *Freiburger Persönlichkeitsinventar – revidierte Fassung* (FPI-R, Fahrenberg, Hampel & Selg, 1984) eher reaktantes Antwortverhalten hervorruft als das vierkategoriale Antwortformat des *Trierer Persönlichkeitsfragebogens* (TPF; Becker, 1989). Zwar werden die gefundenen Reihenfolgeneffekte auf die unterschiedlichen Antwortformate zurückgeführt, dennoch sind die Effekte mit Blick auf die vorliegende Studie auch in anderer Hinsicht relevant. Die Bearbeitung der Leistungstests vor der Beantwortung der Persönlichkeitsfragebögen führte in einigen Skalen des FPI-R zu Testreihenfolgeneffekten. Diese Effekte können nicht nur auf die unterschiedlichen Antwortformate, sondern auch auf die geringe kognitive Anforderung der Leistungstests zurückgeführt werden. Leistungstests mit geringer kognitiver

Anforderung an die Testpersonen können anscheinend das Antwortverhalten bei Persönlichkeitsfragebögen beeinflussen.

9.3.3. Testreihenfolgeeffekte bei Persönlichkeitsfragebögen und OPTs

Baldinger (2006) führte eine Studie zum Einfluss sozial erwünschten Antwortverhaltens bei Persönlichkeitsfragebögen und OPTs durch. Dabei wurden Personen aus einem realen Auswahlprozess mit Personen aus einer Potenzialanalyse verglichen. Vorgegeben wurden der objektive Persönlichkeitstest AHA, der auch in der vorliegenden Studie Verwendung fand, und der Berufseignungsfragebogen *Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung* (BIP; Hossiep, Paschen & Mühlhaus, 2003). Zwei unterschiedliche Vorgabereihenfolgen der Verfahren wurden realisiert. Bei den Analysen konnte in lediglich einem Kennwert ein Effekt der Vorgabereihenfolge festgestellt werden. Betroffen waren jene Personen, die zuerst die AHA und dann den BIP bearbeiteten. Sie zeigten im Kennwert Entschlussfreudigkeit des AHA signifikant niedrigere Werte, waren also weniger entschlussfreudig, wenn die Testung mit den AHA anfang. Dieser Effekt wird mit der weiter fortgeschrittenen Testdauer und der damit zusammenhängenden Gewöhnung an die Testsituation nach der vorausgehenden halbstündigen Bearbeitung des BIPs begründet. Khorramdel (2010) vermutet in diesem Zusammenhang, dass mit fortschreitender Testdauer eine Steigerung der Arbeitsgeschwindigkeit einhergeht.

Darüber hinaus ist zu berichten, dass Unterschiede zwischen der Gruppe von Personen, die die Verfahren in einem realen Auswahlprozess bearbeiteten, und denen, die auf freiwilliger Basis an einer Potenzialanalyse teilnahmen, gefunden wurden. In fast allen Kennwerten des BIP, jedoch in keinem Kennwert der AHA, zeigten sich signifikante Unterschiede zwischen diesen beiden Gruppen. Es konnten also bedeutsame Unterschiede zwischen freiwilligen und nicht freiwilligen Stichproben bei Persönlichkeitsfragebögen festgestellt werden, die in diesem Umfang bei dem verwendeten OPT nicht auftraten.

Eine weitere Studie mit ganz ähnlicher Testauswahl wurde von Janschek (2011) durchgeführt. Diese Arbeit beschäftigte sich mit dem Vergleich der Verfälschbarkeit von Persönlichkeitsfragebögen und OPTs. Bewerbern und Mitarbeitern eines

Unternehmens wurden in zwei unterschiedliche Vorgabereihenfolgen die Verfahren BIP, AHA und NEO-PI-R (*NEO-Persönlichkeitsinventar nach Costa und McCrae*; Ostendorf & Angleitner, 2004) vorgegeben. Die Reihenfolgenvariation war hier im Versuchsplan zur Kontrolle von Reihenfolgeneffekten vorgesehen. Es zeigten sich Reihenfolgeneffekte im Kennwert Entschlussfreudigkeit der AHA und bei einem Kennwert des BIPs. Beim NEO-PI-R wurde in einem Kennwert ein Interaktionseffekt der Vorgabereihenfolge mit der experimentellen Gruppenzugehörigkeit festgestellt.

9.3.4. Testreihenfolgeneffekte bei Leistungstests

Testreihenfolgeneffekte bei ausschließlicher Verwendung von Leistungstests wurden von Eiselt (1991) bei Variation der Vorgabereihenfolge in einer computerisierten Testbatterie untersucht. Zwei Leistungstests wurden in jeweils unterschiedlicher Reihenfolge vorgegeben, dabei konnte ein signifikanter Effekt der Vorgabereihenfolge in einigen Kennwerten festgestellt werden. Beide Gruppen zeigten dasselbe Antwortverhalten. Eine gesteigerte Arbeitsgeschwindigkeit bei geringerer Anzahl vollständig bearbeitete Aufgaben gegen Ende der Testung ließen einen Ermüdungseffekt vermuten. Die abfallende Leistungsfähigkeit im Testverlauf wird vom Autor auf mangelnde Leistungsbereitschaft der Teilnehmer und somit auf Motivationsprobleme zurückgeführt. Das Müdigkeitsempfinden der Testteilnehmer wurde ebenfalls erhoben. Wie im Abschnitt II.8. bereits dargelegt, führt Ermüdung nur in seltenen Fällen tatsächlich zu einem Abfall der Leistungsfähigkeit. Der Autor konnte das Auftreten von Reihenfolgeneffekten lediglich mit der subjektiven Motivation und nicht mit dem subjektiven Müdigkeitsempfinden der Teilnehmer in Verbindung bringen.

Im Zuge der gemeinsamen Normierung eines Intelligenz- und eines Gedächtnisleistungstests wurden die Normierungsstichproben beider Verfahren von Zhu und Tulskey (2000) auf Effekte der Vorgabereihenfolge untersucht. Es ist ein gängiges Vorgehen bei Normierungsprojekten, mehrere Verfahren gleichzeitig vorzugeben, wobei die Reihenfolge aus Gründen effizienter paralleler Testungen variiert werden kann. Problematisch ist dieses Vorgehen, wenn die Bearbeitung des einen Verfahrens die Ergebnisse des anderen beeinflusst. Die Verlässlichkeit der

Normierungsstichprobe wird so verwässert, und konfundierende Testreihenfolgeeffekte gefährden die interne Validität der Verfahren.

Anhand der Normierungsstichprobe der *Wechsler Adult Intelligence Scale – Third Edition* (WAIS-III; Wechsler, 1997) und der *Wechsler Memory Scale – Third Edition* (WMS-III; Wechsler, 1997) konnten Testreihenfolgeeffekte untersucht werden, da beide Verfahren gemeinsam normiert wurden. Die mehr als 1200 Personen umfassenden Normierungsstichprobe setzt sich aus zwei Gruppen zusammen, die eine unterschiedliche Testreihenfolge bearbeiteten. Jeweils etwa die Hälfte der Personen bearbeitet zunächst den WAIS-III und dann den WMS-III und umgekehrt. Die Autoren vermuteten, dass *carry-over*-Effekte aufgrund der inhaltlichen Ähnlichkeit der Testbatterien auftreten würden. In Anbetracht der beträchtlichen Vorgabedauer von etwa 5 Stunden wurden ebenfalls Ermüdungseffekte erwartet. Es konnte jedoch kein Haupteffekt der Testreihenfolge gefunden werden, allerdings gab es signifikante Interaktionseffekte zwischen der Testreihenfolge und den vorgegebenen Subtests. Je nach Inhalt wurden bei einigen Subtests in Abhängigkeit der Vorgabereihenfolge signifikant unterschiedliche Werte festgestellt. Allerdings war es den Autoren nicht möglich Schlüsse aus diesen Interaktionseffekten zu ziehen. Es kann jedoch festgehalten werden, dass vor allem jene Subtests betroffen waren, die Gedächtnis- und Aufmerksamkeitsleistungen erforderten. Ein Muster oder ein erkennbarer Zusammenhang hinsichtlich der teilweise beliebig variierenden Leistungssteigerung und Leistungsminderungen zeichnete sich jedoch nicht ab. Die Autoren merkten an, dass die Ergebnisse inkonsistent seien und verwiesen auf die geringen Effektstärken. Sie erklärten die vereinzelt signifikanten Werte mit statistischen Artefakten, die bei Signifikanztestungen an großen Stichproben auftreten können. Im Hinblick auf die vorliegende Arbeit ist hervorzuheben, dass die Vorgabedauer der Verfahren in der Studie von Zhu und Tulsky (2000) bei etwa 5 Stunden lag. Damit bewegt sich die Gesamtbearbeitungszeit ungefähr in den Bereichen der in der vorliegenden Arbeit verwendeten Testbatterie.

In einer zweiten Studie (Tulsky & Zhu, 2000) wurden an einem Teil der von Zhu und Tulsky (2000) verwendeten Normierungsstichprobe erneute Analysen durchgeführt. Es wurde untersucht, ob die Testleistung in einem Subtest zum Arbeitsgedächtnis (*Buchstaben-Zahlen-Folge*) vom Fortschritt der Testdauer beeinflusst wird. Drei Versuchsgruppen wurden gebildet, wobei die erste den Subtest nach etwa einer Stunde, die zweite nach etwa zweieinhalb Stunden und die dritte Gruppe nach etwa

fünf Stunden Testdauer absolvierte. Es wurde angenommen, dass signifikante Gruppenunterschiede durch Ermüdungseffekte mit einem tendenziellen Leistungsverfall bei steigender Testdauer verursacht würden. Es konnte jedoch kein solcher Effekt gefunden werden. Die Testleistung war also bei unterschiedlichen Testreihenfolgen nicht durch die verstrichene Bearbeitungszeit beeinflusst. Die Testpersonen zeigten keine Ermüdungseffekte, tatsächlich war die Testleistung in allen Konditionen nahezu gleich. Um der Annahme nachzugehen, dass ältere Testpersonen eher Ermüdungseffekte zeigen als jüngere Testpersonen, wurde der Datensatz in einer weiteren Analyse in drei Altersgruppen unterteilt. Auch in dieser Analyse konnten keine signifikanten Unterschiede gefunden werden. Weder die Bearbeitungszeit noch das Alter hatten somit Einfluss auf die Testleistung bei unterschiedlicher Testreihenfolge.

Ryan, Glass, Hinds und Brown (2010) variierte die Vorgabereihenfolge des *Boston Naming Tests* (BNT; Kaplan, Goodglass & Weintraub, 1983) und des *Test of Memory Malinger* (TOMM; Tombaugh, 1996), um einen möglichen Einfluss der Testreihenfolge auf Gedächtnisleistungen abzuklären. Eine nicht klinische Stichprobe aus Universitätsstudenten wurde für diese Studie herangezogen. Der TOMM findet bei simulationsdiagnostischen Fragestellungen zur Feststellung tatsächlicher oder vorgetäuschter Gedächtnisbeeinträchtigungen Verwendung. Der BNT ist ein Screening Tool zur Abklärung neuropsychologischer Hirnschädigungen. Zusätzlich wurde in einer Versuchsgruppe ein visueller Test (*Bilder Vervollständigen* aus dem WAIS-III) zwischen die beiden Testteile des TOMM eingefügt. Es konnten bei Variation der Vorgabereihenfolge keine Effekte auf die Gedächtnisleistung festgestellt werden. Lediglich wenn der BNT zu Beginn vorgegeben wurde, konnte in einem Kennwert eine signifikante Veränderung festgestellt werden. Allerdings liefern die Autoren hierfür keine inhaltliche Erklärung, bezeichnen diese Abweichung als unerheblich und raten lediglich von einer Vorgabe in dieser Reihenfolge ab.

Llorente, Sines, Rozelle, Turcich und Casatta (2000) führen mehrere Studien an, in denen bei diversen Verfahren - zumeist Leistungstests aus dem klinischen Bereich - Vorgabereihenfolgen variiert wurden. Nicht signifikante Ergebnisse (Cassel, Johnson & Burns, 1962; Neuger, O'Leary, Fishburne, Barth, Berent, Giordani & Boll, 1981), aber auch signifikante Reihenfolgeeffekte werden berichtet (Brandon & Chavez, 1985; Franzen, Smith, Paul & MacInnes, 1993). Eine einheitliche Schlussfolgerung aus den genannten Arbeiten liefern Llorente et al. (2000) allerdings nicht - es wird lediglich

vermutet, dass Reihenfolgeeffekte vom Grad der Gemeinsamkeit der verwendeten Verfahren abhängig seien. Je höher der Anteil geteilter Varianz zweier Verfahren ist, umso anfälliger sind sie für eine gegenseitige Beeinflussung. So sind beispielsweise die von Faust und Hollingsworth (1991) gemeinsam variierten Intelligenz- und Sprachfähigkeitstests konzeptuell weit voneinander entfernt, wodurch das Ausbleiben von Reihenfolgeeffekten erklärt werden könnte. In einer eigenen Studie untersuchten Llorente et al. (2000) Testreihenfolgeeffekte anhand zweier Sprachfähigkeitstest wobei eine Beeinflussung in lediglich einer Vorgabereihenfolge angenommen wurde. Diese Annahme konnte bestätigt werden und wird als Beleg für starke Lerneffekte bei Leistungstests aufgrund inhaltlicher Ähnlichkeit der Verfahren ausgelegt. Dabei diente der eine Test quasi als Übungsphase für den folgenden, und so erhöhte sich verständlicherweise die Leistung im darauf folgenden Verfahren. Dieser Effekt lag aufgrund inhaltlicher Aspekte der verwendeten Verfahren lediglich in eine Richtung vor, da der Lerneffekt nur in einer Reihenfolge effektiv gegeben war und zu einer Vereinfachung des Folgetests geführt hatte.

10. Testreihenfolgeeffekte in der Personalauswahl

Im folgenden Abschnitt werden jene beiden Studien vorgestellt, die maßgeblich zur Erforschung von Testreihenfolgeeffekten in der Personalauswahl beigetragen haben. Es handelt sich um zwei Experimente, die der vorliegenden Studie vorangegangen sind. Die in diesen Studien gewonnenen Erkenntnisse und das verwendete Forschungsdesign sind die Ausgangspunkte der vorliegenden Arbeit. Im weiteren Verlauf wird immer wieder Bezug auf diese beiden Studien genommen, beide sind Teil der Dissertation von Khorramdel (2010). Die erste Studie wird hier als Experiment 1, die zweite daran anschließende und bisher unpublizierte Studie als Experiment 2 bezeichnet. Experiment 3 bezeichnet die vorliegende Arbeit, sie stellt die aktuellste Folgestudie dar.

Allen drei Studien ist gemein, dass sie jeweils in realen Auswahl-situationen durchgeführt werden konnten. Die herangezogenen Stichproben zeichnen sich dadurch aus, dass für die Teilnehmer berufliche und persönliche Konsequenzen aus den Ergebnissen der Testungen entstehen konnten. Dieser Umstand garantierte eine hohe Motivation der Teilnehmer.

Für die Testungen wurde eine umfangreiche Testbatterie bestehend aus Leistungstests und OPTs, vorgegeben, wodurch einerseits die Erprobung dieser Verfahren in der Praxis ermöglicht wurde und andererseits Testreihenfolgeeffekte bei Kombination dieser beiden Verfahrensgattungen in einem wissenschaftlichen Experiment untersucht werden konnten.

Zwar steht primär die Erforschung von Testreihenfolgeeffekten im Fokus aller drei Studien, sie unterscheiden sich jedoch leicht in der Ausrichtung ihrer Forschungsfragen. Es werden unterschiedliche Vermutungen über die Ursachen von Testreihenfolgeeffekten aufgestellt und durch systematische Variation über alle drei Studien hinweg experimentell untersucht. Die Ergebnisse und Interpretationsansätze werden im Folgenden dargestellt, sie sind für das Verständnis der hier vorliegenden Arbeit von großer Bedeutung.

10.1. Experiment 1: Khorramdel & Frebort (2011)

Khorramdel und Frebort (2011) führten in Kooperation mit einem Unternehmen aus der Automobilzuliefererbranche eine Studie zu Testreihenfolgeeffekten durch. Die Stichprobe umfasste ausschließlich Personen aus gehobenen Managementpositionen des Unternehmens, die an einer Evaluation ihrer kognitiven Fähigkeiten und tätigkeitsrelevanter Persönlichkeitseigenschaften teilnahmen. Es sollte anhand berufsbezogener Leistungs- und Persönlichkeitsvariablen festgestellt werden, ob sie für ihre derzeitige Position im Unternehmen geeignet sind. Im Vorfeld wurde ein Anforderungsprofil für Führungskräfte erstellt. Darauf aufbauend, erfolgte die Zusammenstellung einer Testbatterie zur adäquaten Erfassung dieser Anforderungen. Im Hinblick auf die bereits mehrfach angedeuteten Probleme bei der Vorgabe von Persönlichkeitsfragebögen in Auswahlverfahren (Kubinger, 2002) wurden hier ausschließlich kognitive Leistungstests und OPTs herangezogen.

Zum Einsatz kamen die kognitiven Leistungstests *Adaptiver Matrizentest* (AMT; Hornke, Etzel & Rettig, 2007) sowie der *Intelligenz-Struktur-Test 2000 - revidierte Fassung* (IST 2000-R; Liepmann, Beauducel, Brocke & Amthauer, 2007). Es wurden dabei nur die Subtests zur verbalen und numerischen Intelligenz des IST 2000-R vorgegeben. Zur Gruppe der vorgegebenen OPTs zählten die Arbeitshaltungen (AHA; Kubinger & Ebenhö, 2007), das *Belastbarkeits-Assessment BaCO-D* (BaCO-D;

Ortner, Kubinger, Schrott, Radinger & Litzenberger, 2007) sowie ein *Simulationstest zur Erfassung des Entscheidungsverhaltens* (ILICA; Möseneder & Ebenhöf, 1996). Die verwendeten Verfahren wurden jeweils anhand ihrer Verfahrensgattung zu Blöcken zusammengefasst vorgegeben. Die beiden Versuchsbedingungen unterschieden sich dann lediglich durch die Vorgabereihenfolge der Blöcke, während die Reihenfolge der Verfahren innerhalb der Blöcke konstant gehalten wurde.

Die Stichprobe umfasste insgesamt 66 Teilnehmer im Altersbereich von 26 bis 54 Jahren mit mehrjähriger Führungserfahrung. Für die Teilnehmer waren die Ergebnisse der Testungen von hoher persönlicher Relevanz, da eine negative Evaluation einen potentiellen Verlust der aktuellen Position im Unternehmen bedeuten konnte. Auch wenn diese harten Konsequenzen im Rahmen einer wissenschaftlichen Studie als durchaus ethisch problematisch angesehen werden können, muss darauf verwiesen werden, dass nach hohen wissenschaftlichen Standards und mit bewährten psychologisch diagnostischen Verfahren durchgeführte Erhebungen anderen üblicherweise im Personalwesen anzutreffenden Vorgehensweisen in punkto Fairness weit überlegen sind.

Das wissenschaftliche Hauptinteresse dieser Studie bestand darin, die Folgen einer Testreihenfolgenvariation zu erforschen, die durch zwei unterschiedliche Vorgabereihenfolgen bei zwei Versuchsgruppen realisiert wurde. Die Studie basiert auf der Annahme, dass die auf Itemebene bereits gut erforschten Reihenfolgeeffekte auch bei der Variation ganzer Verfahrensblöcke in einer Testbatterie auftreten können. Nach Fazio, Powell und Herr (1983) sowie La Rue und Olejnik (1980) soll die Vorgabe mehrerer unterschiedlicher Verfahren innerhalb einer Testbatterie die Leistung der Testpersonen im Verlauf der Bearbeitung beeinflussen können. Es wurde das Auftreten von sich auf die Performanz der Testpersonen auswirkenden Kontexteffekten hypothetisiert. Erwartet wurden hauptsächlich die im Abschnitt II.8. angeführten *carry-over*-, *Priming*-, Lern- und Ermüdungseffekte. Anhand von Mittelwertsunterschieden in den Kennwerten der vorgegebenen Verfahren würden sich Effekte der Testreihenfolge feststellen lassen. Das experimentelle Design erlaubt es, Informationen über stichproben- und verfahrensbedingte Umstände zu sammeln, die für Testreihenfolgeeffekte verantwortlich sein können.

Die statistischen Analysen konnten einen signifikanten Haupteffekt der Testreihenfolge aufzeigen. Anschließende univariate Analysen identifizierten die

Kennwerte Entschlussfreudigkeit und Frustrationstoleranz der AHA als Hauptursache für den signifikanten Haupteffekt. Zwischen den Versuchsgruppen wichen nur diese beiden Kennwerte signifikant voneinander ab und waren in der Gruppe, die zuerst die Leistungstests und anschließend die OPTs bearbeitete, signifikant höher beziehungsweise niedriger ausgeprägt. Im Detail zeigten jene Teilnehmer, die bereits ungefähr zweieinhalb Stunden mit der Testung verbracht hatten, da sie die Testung mit den Leistungstests begonnen hatten, bei der Bearbeitung der AHA eine höhere Entschlussfreudigkeit und eine niedrigere Frustrationstoleranz. Sie reagierten schneller auf die vorgegebenen Items und zeigten bei (falsch) negativem Feedback über ihre Leistung eine geringere Frustrationstoleranz. In Übereinstimmung mit den Ergebnissen von Baldinger (2006) erwies sich der Kennwert Entschlussfreudigkeit des AHA also erneut im Zusammenhang mit Reihenfolgevariationen als auffällig. Die Autoren vermuteten also, dass dieser Kennwert möglicherweise unabhängig von den vorgegebenen Verfahren generell anfällig für Reihenfolgeeffekte sei. Eine weitere Parallele zu Baldinger (2006) besteht darin, dass die Stichprobe in beiden Studien durch ein hohes Leistungsniveau charakterisiert war. In Verbindung mit der geringen kognitiven Anforderung, die das Itemmaterial des AHA an die Testpersonen stellt, wirkt sich dies nach Meinung der Autoren vermutlich begünstigend auf das Auftreten von Testreihenfolgeeffekten aus.

Die signifikanten Unterschiede in den Kennwerten können inhaltlich als Abnahme der Leistungsfähigkeit interpretiert werden. Sie zeigten sich lediglich in jener Gruppe, in der der Testverlauf bereits weiter fortgeschritten war. Somit wurden vor allem Ermüdungseffekte als mögliche Ursache der Testreihenfolgeeffekte vermutet. Die erhöhten Werte im Kennwert Entschlussfreudigkeit wurden dahingehend interpretiert, dass die bereits verstrichene Bearbeitungszeit die Teilnehmer ein Ende der Testsituation herbeisehnen lasse und diese durch verringerte Antwortzeiten herbeizuführen gesucht werde. Die intensive zeitliche und kognitive Beanspruchung durch die vorangegangene Bearbeitung von Leistungstest verstärkte diesen Effekt zusätzlich. Die Bearbeitung der vergleichsweise leichten Aufgaben des AHA ermöglichten eine Erhöhung der Arbeitsgeschwindigkeit, was sich wiederum im Kennwert Entschlussfreudigkeit niederschlug.

Bezüglich des Kennwertes Frustrationstoleranz wurden konkrete Überlegungen unter Einbezug von Stichprobencharakteristika angeführt. Die Autoren vermuteten einen Zusammenhang zwischen der Art und Weise, in der Frustration in den AHA über

soziale Vergleiche herbeigeführt wird, und der Persönlichkeitsstruktur der Teilnehmer. Es wurde angenommen, dass Führungskräfte kompetitiven sozialen Vergleichsprozessen eine hohe Bedeutung beimessen und dass sie somit durch die Operationalisierung besonders betroffen sind. Die Stichprobe war durch Personen mit hohem Leistungsniveau charakterisiert, denen Erfolg und Leistung persönlich wichtig sind. Die vorherige Bearbeitung fordernder Leistungstests könnte die frustrierende Wirkung sozialer Vergleiche verstärkt haben. Ergänzend wurde eine Verbindung zwischen dem Auftreten von Reihenfolgeeffekten und der Erfassung von Variablen vermutet, die im Zusammenhang mit Resilienzkonzepten stehen. Dies hatte sich ebenfalls bei Hambros (2002) in Bezug auf Persönlichkeitsfragebögen gezeigt. Frustrationstoleranz steht im Zusammenhang mit Resilienz und *Coping*-Strategien, da eine Abhärtung gegen Frustrationen günstigen Einfluss auf den Umgang mit frustrierenden Situationen hat.

Zusammengefasst deuten die Ergebnisse dieser Studie darauf hin, dass fordernde Leistungstests weniger anfällig für Reihenfolgeeffekte sind, wohingegen OPTs durchaus von der Vorgabereihenfolge betroffen zu sein scheinen. Insbesondere bei wenig anspruchsvollem Itemmaterial und bei Testpersonen mit hohem Leistungsniveau wurden diese Effekte deutlich. Einschränkend wurde angeführt, dass die Stichprobe zu einem überwiegenden Teil aus Männern eines einzigen Unternehmens bestand, die ausschließlich in Positionen des gehobenen Managements tätig waren. Die Generalisierbarkeit der Ergebnisse ist somit eingeschränkt. In zukünftigen Studien sollte daher vor allem eine Erhöhung der Generalisierbarkeit durch weniger spezifische Stichproben erreicht werden und es sollten in besonderem Ausmaß Resilienz- und Belastbarkeitsaspekte untersucht werden, da sich diese wiederholt als anfällig für Reihenfolgeeffekte erwiesen haben.

10.2. Experiment 2: Khorramdel (2010)

Im Zuge der Zusammenarbeit mit dem Unternehmen aus Experiment 1 führte Khorramdel (2010) eine Folgestudie unter Berücksichtigung der Ergebnisse und der Kritikpunkte an Experiment 1 durch. Erneut wurden Mitarbeiter des Unternehmens einer Testung mit einer ähnlichen Testbatterie, bestehend aus kognitiven Leistungstests und OPTs, unterzogen.

Die Ergebnisse aus Experiment 1 sollten repliziert werden, und darüber hinaus sollte erforscht werden, ob die dort gefundenen signifikanten Testreihenfolgeeffekte eher auf die verwendeten Tests oder aber auf Stichprobencharakteristika zurückführbar sind. Angelehnt an die Modelle von Petty und Cacioppo (1986) sowie Schwarz, Hippler und Noelle-Neumann (1992), welche die Motivation und die kognitive Fähigkeit der Testteilnehmer als maßgeblich für die Leistung in einem Test herausstreichen, sollten diese Aspekte genauer untersucht werden.

Die herangezogene Stichprobe umfasste 64 Personen, diese waren ausschließlich Männer im Altersbereich von 22 bis 56 Jahren. Die Testungen wurden wie in Experiment 1 im Rahmen einer Evaluation kognitiver Fähigkeiten und tätigkeitsbezogener Persönlichkeitseigenschaften durchgeführt. Die Teilnehmer wurden ebenfalls hinsichtlich ihrer Eignung für ihre derzeitige Position im Unternehmen beurteilt. Das Potential für eine Beförderung auf eine höhere Ebene sollte abgeklärt werden, allerdings konnte bei negativer Evaluation auch eine Degradierung in eine Position ohne Führungsverantwortung aus der Testung resultieren. Es konnte dabei erneut auf Teilnehmer mit Führungserfahrung zurückgegriffen werden, diese waren jedoch auf niedrigeren Führungsebenen des Unternehmens tätig.

Ein Anforderungsprofil, welches sich größtenteils mit dem aus Experiment 1 deckt, wurde erstellt. Eine leicht veränderte Verfahrensauswahl war dennoch notwendig. Im Unterschied zur Testbatterie in Experiment 1 wurden anstelle des AMTs die *Standard Progressive Matrices* (SPM; Raven, Raven & Court, 2008) vorgegeben. Dabei handelt es sich im Gegensatz zum AMT um einen nicht adaptiven Matrizentest mit einem geringeren Schwierigkeitsgrad. Darüber hinaus wurden die Subtests zur Merkfähigkeit des IST 2000-R in die Testbatterie integriert. Somit wurde zum einen die kognitive Beanspruchung variiert und zum anderen die Merkfähigkeit als zusätzliche Variable erfasst, die sich zuvor bereits als anfällig für Reihenfolgeeffekte erwiesen hatte (Zhu & Tulsky, 2000).

Das experimentelle Design entsprach dem in Experiment 1 verwendeten. Zwei Versuchsgruppen wurden die geblockt variierten Leistungstests und OPTs vorgegeben. Erneut wurde die Primärhypothese geprüft, ob die Testreihenfolge einen Einfluss auf die Testleistung hat. Zusätzliche Sekundärhypothesen über das Auftreten oder Ausbleiben von Testreihenfolgeeffekten wurden für den Fall inkonsistenter Ergebnisse aufgestellt. Testreihenfolgeeffekte könnten in Bezug auf die Modelle von

Petty und Cacioppo (1986) sowie Schwarz, Hippler und Noelle-Neumann (1992) in Abhängigkeit bestimmter Stichprobencharakteristika auftreten. Es wurde also geprüft, ob zwischen den Teilnehmern von Experiment 1 und 2 Unterschiede hinsichtlich kognitiver Fähigkeiten oder der Leistungsmotivation vorlagen.

Die Datenanalyse verlief ebenfalls vergleichbar zu Experiment 1. So wurden zunächst sämtliche erhobenen Kennwerte auf einen Haupteffekt der Testreihenfolge untersucht. Dieser konnte jedoch nicht festgestellt, das Ergebnis aus Experiment 1 somit nicht repliziert werden. Daran anschließende univariate Analysen zeigten, dass die Kennwerte zur Merkfähigkeit des IST 2000-R signifikante Gruppenunterschiede aufwiesen. Die signifikanten Unterschiede in der Merkfähigkeit wurden mit einem leichten Ermüdungseffekt erklärt, da diese lediglich in jener Gruppe signifikant geringer ausfielen, in der die Leistungstests im Anschluss an die OPTs bearbeitet wurden, die Testdauer bei Bearbeitung der Merkfähigkeitstests also bereits weiter fortgeschritten war. Es konnten darüber hinaus keine Testreihenfolgeeffekte festgestellt werden.

Im Hinblick auf die Sekundärhypothesen wurden die Stichproben aus Experiment 1 und 2 hinsichtlich sämtlicher Kennwerte auf signifikante Unterschiede überprüft. Aufgrund der unterschiedlichen Position der Teilnehmer in der Unternehmenshierarchie wurde angenommen, dass die Stichproben durch ein unterschiedliches Bildungsniveau und eine unterschiedlich ausgeprägte Leistungsmotivation charakterisiert seien. Hier konnten signifikante Unterschiede zwischen den beiden Stichproben gefunden werden. Die Stichprobe aus Experiment 1 zeichnete sich durch eine höhere verbale und numerische Intelligenz aus und zeigte außerdem eine höhere Ausprägung im Kennwert Anspruchsniveau bei geringeren Werten in der Zieldiskrepanz. Das logisch schlussfolgernde Denken war ebenfalls unterschiedlich ausgeprägt, allerdings wurden zur Erfassung zwei unterschiedliche Verfahren (AMT und SPM) verwendet, die Ergebnisse sind somit nur eingeschränkt interpretierbar. Darüber hinaus erwiesen sich die Teilnehmer in Experiment 2 als resilienter in Belastungssituationen.

Grundsätzlich konnten in dieser Studie die Ergebnisse aus Experiment 1 nicht repliziert werden, es konnten jedoch sehr wohl neue Erkenntnisse über mögliche Ursachen von Testreihenfolgeeffekten gewonnen werden. Der Vergleich der Stichproben aus Experiment 1 und 2 erlaubte es, weiterführende Annahmen über die Umstände zu machen, unter denen Testreihenfolgeeffekte auftreten. Die Ergebnisse

deuten darauf hin, dass Testreihenfolgeneffekte in Abhängigkeit vom kognitiven Leistungsniveau der Stichprobe auftreten und von motivationalen Aspekten und der Resilienzfähigkeit beeinflusst sind. Darüber hinaus sind auch Aspekte der verwendeten Verfahren relevant. Leichte Aufgaben, gemessen am Potential der Stichprobe, scheinen eher für Testreihenfolgeneffekte anfällig zu sein. Die beiden Matrizentests AMT und SPM, die zwischen beiden Studien variiert wurden, unterscheiden sich hinsichtlich ihrer adaptiven und nicht adaptiven Vorgabe. Wie im Abschnitt II.5. deutlich gemacht wurde, geht adaptives Testen mit einer höheren Schwierigkeit und somit mit einer größeren Beanspruchung der Testperson einher. Die Bearbeitung des AMTs steht somit auch aufgrund der adaptiven Vorgabe im Verdacht, in besonderem Maße frustrierend zu sein. Diese Beanspruchung äußert sich dann in den betreffenden Kennwerten derjenigen Verfahren der Testbatterie, die diese erfassen sollen, beispielsweise in den AHA. Die Testbatterie zeichnet also die durch sie selbst verursachten Testreihenfolgeneffekte auf, sofern die Testreihenfolge dies entsprechend ermöglicht. Wenn die Verfahren in dieser Reihenfolge vorgegeben werden kann es daher zu einer Interaktion von AMT und AHA kommen.

Abschließend wurde aufgrund der inkonsistenten Ergebnisse aus Experiment 1 und 2 von einer Variation der Testreihenfolge abgeraten. Sofern die involvierten Effekte nicht weiter erforscht und verstanden sind, lassen sie sich nicht eindeutig identifizieren und kontrollieren. In der Praxis sollte bei derzeitigem Wissensstand von einer Variation der Testreihenfolge abgesehen werden. Die weiterführende Erforschung von Testreihenfolgeneffekten sollte sich der systematischen Zusammenstellung von Stichprobencharakteristika und Item- sowie Testinhalten widmen. Die bisherigen Ergebnisse deuten auf eine komplexe Interaktion von Stichprobencharakteristika, Verfahrensgattung sowie Inhalt und Modalität der Testung hin.

Einschränkend wird noch darauf hingewiesen, dass die Stichproben in den Experimenten 1 und 2 in hohem Maße spezifisch sind und daher eine Generalisierung der Ergebnisse erschwert wird. Zukünftige Studien könnten sich der Bedeutung adaptiver und nicht-adaptiver Tests sowie berufsfeldspezifischer Aspekte widmen. Auch geschlechtsspezifische Effekte sind denkbar, in Experiment 1 und 2 haben zum überwiegenden Teil Männer an den Testungen teilgenommen. Konkrete Aussagen sind in dieser Hinsicht also bisher ebenfalls nicht möglich.

Überleitung

Die im vorangegangenen Abschnitt angeführten theoretischen Überlegungen und empirischen Erkenntnisse zeugen zwar davon, dass es sich um einen seit Jahren aktiven und sich entwickelnden Forschungsbereich handelt, es wird jedoch auch der wenig zufriedenstellende Stand des Wissen um Reihenfolgeeffekte deutlich. Die teilweise inkonsistenten Forschungsergebnisse zeugen von der Notwendigkeit, sich mit dem Forschungsbereich Testreihenfolgeeffekte vertiefend auseinanderzusetzen - vor allem, wenn man die praktische Relevanz der Erkenntnisse bei Personalauswahlprozessen im Unternehmenskontext bedenkt. Insbesondere die Experimente 1 und 2 geben eindeutige Ansatzpunkte für weiterführende Studien. Sie geben Einblick in die Komplexität und Evidenz von Testreihenfolgeeffekten. Die theoretischen Überlegungen, die aus diesen Forschungsergebnissen abgeleitet wurden, gilt es nun auf experimentellem Wege zu klären. Im Anschluss an diese beiden Studien wird in logischer Folge den Fragen nach der Bedeutsamkeit der Verfahrensauswahl und der Stichprobenspezifität für Testreihenfolgeeffekte weiterführend nachgegangen. An Kritikpunkte wird angeknüpft, und es wird der Versuch unternommen, einige der bisherigen Einschränkungen zu überwinden und dabei mit neuen Erkenntnissen aufzuwarten. Zu diesem Zweck wird im empirischen Teil dieser Arbeit Experiment 3 dargestellt und von den gewonnenen Ergebnissen berichtet. Diese sollen in die bisherigen Erkenntnisse integriert werden, um so das Verständnis für Testreihenfolgeeffekte zu erweitern.

III. Empirischer Teil

Ausgehend von den in Teil II ausgeführten theoretischen Überlegungen wurde der Forschungsansatz von Khorramdel und Frebort (2011) und Khorramdel (2010) fortgeführt. Die Hypothesen aus Experiment 1 und 2 wurden für die vorliegende Arbeit angepasst und erneut in einem vergleichbaren experimentellen Design getestet. Im speziellen wurde der Aspekt der Testauswahl aufgegriffen und untersucht. Die Verfahrensauswahl wurde in Folge der durch die Vorgängerstudien gewonnenen Erkenntnisse erneut systematisch variiert. In der vorliegenden Studie wurde wie in Experiment 1 der AMT anstelle des in Experiment 2 verwendeten Matrizentests SPM vorgegeben. Es sollte so der Einfluss dieser beiden Matrizentests untersucht werden, um der Frage nachzugehen, ob der AMT im Gegensatz zu den SPM für Reihenfolgeeffekte verantwortlich ist. Dafür war es notwendig, eine Stichprobe heranzuziehen, die mit jener aus Experiment 2 vergleichbar ist, um den alleinigen Einfluss der Matrizentests in der Testreihenfolge untersuchen zu können. Es ist letztlich nicht gelungen eine in jeder Beziehung zu Experiment 2 vergleichbare Stichprobe heranzuziehen, weshalb die studienübergreifende Aussagekraft bezüglich der verwendeten Matrizentests eingeschränkt ist. Ursachen und Konsequenzen dieses Umstandes werden in den folgenden Abschnitten dargestellt und diskutiert.

Die Studie konnte im Rahmen des Auswahlprozesses zu einem Trainee-Programm der Österreichischen Bundesbahnen (ÖBB) realisiert werden. Die Integration der Testungen in einen realen Auswahlprozess garantierte, wie in den Vorgängerstudien, eine hohe Motivation und Leistungsbereitschaft der Testpersonen. Wie bereits in Abschnitt II.7. dargestellt, sind Teilnehmer, denen relevante Konsequenzen aus Testungen entstehen, wesentlich motivierter, und die Testergebnisse dementsprechend realistischer in Bezug auf die wahre Fähigkeit der Testpersonen und somit verlässlicher zu interpretieren, als bei freiwilligen Testteilnehmern. Darüber hinaus ermöglichte die Kooperation mit den ÖBB eine Durchführung der Studie unter für dieses Experiment ausgezeichneten Bedingungen. Die ÖBB verfügen über eine hervorragende Testinfrastruktur, die für diese Studie genutzt werden konnte. Die Forschungsfragen, die daraus abgeleiteten Hypothesen, die genaue Versuchsplanung und eine umfassende Beschreibung der Stichprobe im Hinblick auf die Vergleichbarkeit zu den Vorgängerstudien sind im Folgenden ausführlich beschrieben.

1. Ziele der Untersuchung

Insgesamt wurden drei Forschungsfragen formuliert. Die erste Forschungsfrage bezieht sich auf das generelle Auftreten oder Ausbleiben von Testreihenfolgeeffekten in umfangreichen Testbatterien. Dieser Aspekt wird in dieser Studie durch eine Variation der Vorgabereihenfolge zwischen den Versuchsgruppen untersucht und wurde inferenzstatistisch analysiert.

Die zweite Forschungsfrage beschäftigt sich im Detail mit der erneuten systematischen Variation der Verfahrensauswahl unter Berücksichtigung der Erkenntnisse aus den Studien von Khorramdel und Frebort (2011) sowie Khorramdel (2010), d.h. Experiment 1 und 2. Dabei sollte versucht werden, Experiment 2 zu replizieren, jedoch unter Verwendung des AMTs anstelle der SPM, um herauszufinden, ob der Matrizentest für die in Experiment 1 gefundenen Testreihenfolgeeffekte verantwortlich ist.

Die dritte Forschungsfrage fokussiert auf die Bedeutung von Stichprobencharakteristika. Es konnte in der vorliegenden Studie zwar erneut in einem realen Auswahlprozess eine hoch motivierte Stichprobe getestet werden, diese war jedoch hinsichtlich ihres kognitiven Leistungsniveaus, ihres Alters, ihrer Berufs- und Führungserfahrung sowie weiterer Eigenschaften anders charakterisiert als die in Experiment 1 und 2 herangezogenen Stichproben. Diese dritte Forschungsfrage resultiert somit aus der in Abschnitt III.2.4. dargestellten, nicht gelungenen Replikation von Experiment 2 in Bezug auf die Vergleichbarkeit der Stichprobencharakteristika. Es muss hierfür erarbeitet werden, inwiefern die Stichproben vergleichbar sind und worin sie sich unterscheiden. Die Konsequenzen der abweichenden Stichprobencharakteristika werden in Abschnitt III.6. kritisch diskutiert.

1.1. Hypothesen

Bezüglich der ersten Forschungsfrage wird für diese Studie folgendes Hypothesenpaar formuliert und anhand der erhobenen Daten geprüft. Die Hypothesen 1 und 2 bilden dabei ein komplementäres Paar und sollen untersuchen, ob die jeweilige Vorgabe eines Testblockes Einfluss auf die Bearbeitung des darauf folgenden Testblockes hat. Die Hypothesen 1 und 2 sind wie folgt formuliert:

Hypothese 1 Reihenfolgeeffekte LT → OPT:

H1.0: Die Bearbeitung kognitiver Leistungstests beeinflusst die Bearbeitung darauf folgender Objektiver Persönlichkeitstests nicht.

H1.1: Die Bearbeitung kognitiver Leistungstests beeinflusst die Bearbeitung darauf folgender Objektiver Persönlichkeitstests.

Hypothese 2 Reihenfolgeeffekte OPT → LT:

H2.0: Die Bearbeitung Objektiver Persönlichkeitstests beeinflusst die Bearbeitung darauf folgender kognitiver Leistungstests nicht.

H2.1: Die Bearbeitung Objektiver Persönlichkeitstests beeinflusst die Bearbeitung darauf folgender kognitiver Leistungstests.

Zur Klärung der zweiten Forschungsfrage wird eine weitere Hypothese formuliert, die unter Berücksichtigung der Ergebnisse aus Experiment 2 Aufschluss darüber gibt, ob der AMT für die Testreihenfolgeeffekte aus Experiment 1 verantwortlich gemacht werden kann. Für sich genommen behandelt diese Hypothese, ob unter Verwendung des AMTs in der vorliegenden Studie ein Testreihenfolgeeffekt auftritt oder nicht.

Hypothese 3 Testreihenfolgeeffekte bei Vorgabe des AMTs:

H3.0: Die Vorgabe des AMTs hat keinen Einfluss auf das Auftreten von Testreihenfolgeeffekten.

H3.1: Die Vorgabe des AMTs hat einen Einfluss auf das Auftreten von Testreihenfolgeeffekten.

Zur dritten Forschungsfrage ist Folgendes zu beachten: Im Falle einer nicht gewährleisteten Vergleichbarkeit der Stichproben der Experimente 2 und 3, müssen weiterführende Überlegungen angestellt werden, die auf den Unterschieden und Gemeinsamkeiten der in allen drei Studien herangezogenen Stichproben aufbauen. Wie auch bei Khorramdel und Frebort (2011) sollen Annahmen über Stichprobencharakteristika herausgearbeitet werden, die für Testreihenfolgeeffekte relevant und für weiterführende Studien interessant sein können. Diese Fragestellung kann lediglich inhaltlich im Rahmen der Diskussion der vorliegenden Arbeit in Abschnitt III.6. behandelt werden, weshalb hierzu auch keine konkreten Hypothesen formuliert werden.

2. Methode

Zur Klärung der durch die Hypothesen formulierten Fragestellungen wurde ein Experiment durchgeführt. Das Versuchsdesign ermöglicht die Klärung der ersten Forschungsfrage, operationalisiert durch die Hypothesen 1 und 2, mittels inferenzstatistischer Analysen. Die zweite Forschungsfrage kann anhand von Hypothese 3 bei gegebener Vergleichbarkeit der Stichproben aus Experiment 2 und 3 zusammen mit den Erkenntnissen der Vorgängerstudien beantwortet werden. Treten in der hier vorliegenden Studie wie in Experiment 1 Testreihenfolgeeffekte auf, können diese auf die Verwendung des AMTs anstelle des SPMs aus Experiment 2 zurückgeführt werden. Damit wäre der AMT für Testreihenfolgeeffekte verantwortlich.

2.1. Untersuchungsplan und Versuchsdesign

In einem Zwei-Gruppen-Versuchsdesign wurden den Teilnehmern zwei unterschiedliche Testreihenfolgen vorgegeben. Das Versuchsdesign entspricht jenem in Experiment 1 und 2. Die verwendeten Verfahren wurden in zwei Blöcke jeweils zu kognitiven Leistungstests (Block LT) und objektiven Persönlichkeitstests (Block OPT) zusammengefasst. Die Testreihenfolge innerhalb dieser Blöcke wurde nicht variiert, die Blöcke selbst waren jedoch in ihrer Reihenfolge zwischen den Gruppen vertauscht. Gruppe 1 bearbeitete zunächst die kognitiven Leistungstests und darauffolgend die objektiven Persönlichkeitstests (Gruppe 1: LT → OPT), während Gruppe 2 die Blöcke LT und OPT in umgekehrter Reihenfolge bearbeitete, so dass erst die objektiven Persönlichkeitstests und danach die kognitiven Leistungstests vorgegeben wurden (Gruppe 2: OPT → LT). *Abbildung 1* stellt das Versuchsdesign grafisch dar.

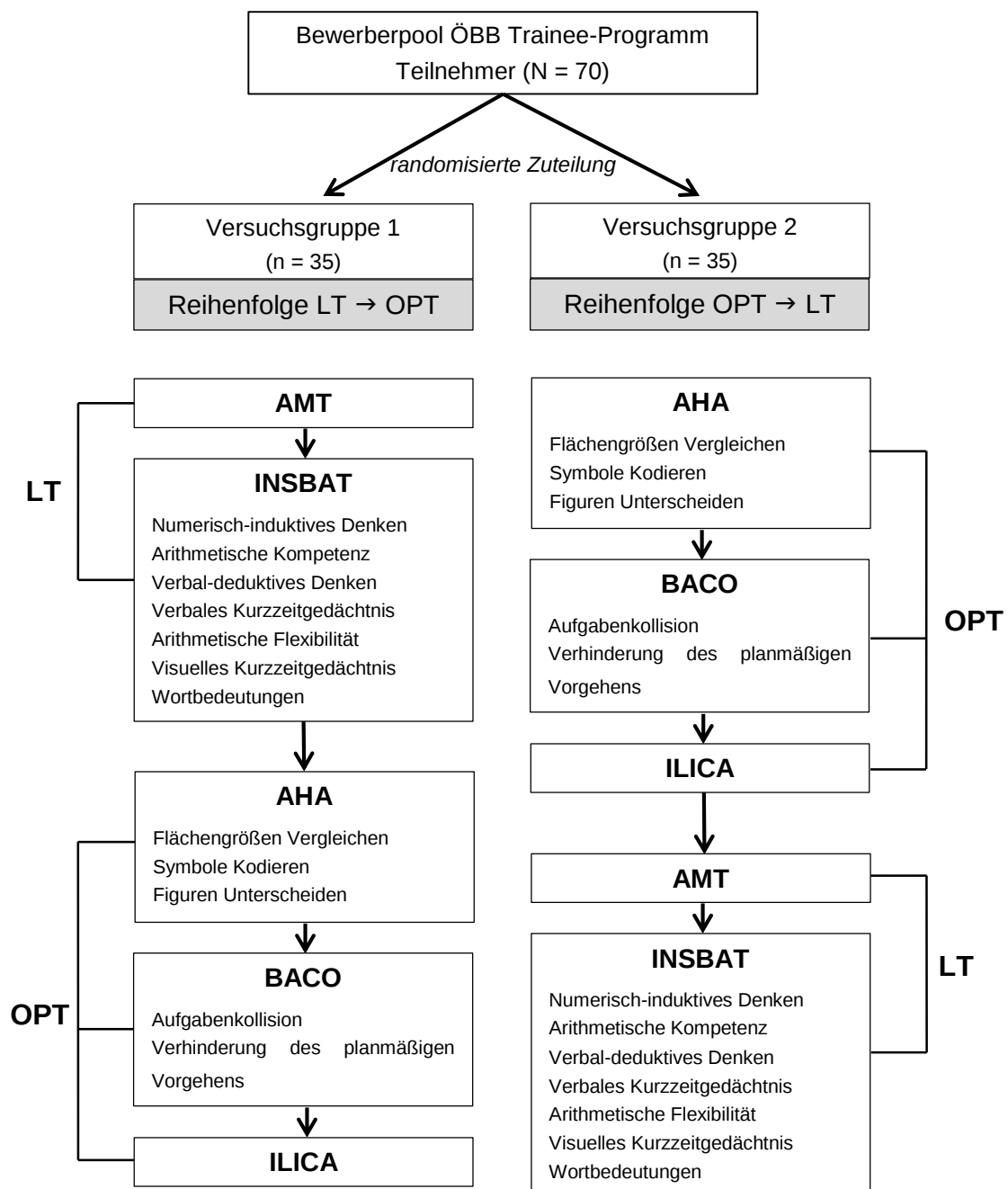


Abbildung 1: Versuchsdesign – LT = Leistungstests, OPT = Objektive Persönlichkeitstests

2.2. Erhebungsinstrumente

Die in Experiment 1 und 2 vorgegebenen Verfahren wurden anhand eines im Vorfeld erstellten Anforderungsprofils ausgewählt. Das Anforderungsprofil verlangte die Qualifikation zur Führungskraft sowie die generelle Eignung für

Führungspositionen in einem Unternehmen aus der Automobilzuliefererbranche. Es umfasst Anforderungen an die kognitive Leistungsfähigkeit sowie tätigkeitsrelevante Persönlichkeitsaspekte. Im Sinne der Vergleichbarkeit der Experimente 1, 2 und 3 wurden weitgehend dieselben Verfahren verwendet. Da es sich in der vorliegenden Studie um eine ähnliche Stichprobe und um ein vergleichbares Anforderungsprofil handelte (zu besetzende Stellen in einem Trainee-Programm für angehende Führungskräfte), war auch in Abstimmung mit den Verantwortlichen der ÖBB die Verwendung der gleichen Testbatterie plausibel.

Mit Ausnahme des ILICAs sind sämtliche in dieser Studie verwendeten Verfahren im Verfahrensinventar der Schuhfried GmbH verfügbar. Die Verfahren werden innerhalb des Wiener Testsystems (WTS) vorgegeben. Das WTS ist ein Tool zur computergestützten psychologischen Diagnostik. Es handelt sich dabei um ein Computerprogramm mit grafischer Benutzeroberfläche, welches die Administration der Vorgabe einer Vielzahl computerisierter Verfahren ermöglicht. Innerhalb dieses Systems können die Verfahrensauswahl und die Vorgabereihenfolge beliebig variiert werden. Dabei können diverse spezifische Einstellungsmöglichkeiten wie etwa gewünschte Reliabilitäten bei adaptiven Verfahren berücksichtigt werden. Das WTS bietet eine umfangreiche Sammlung psychologisch diagnostischer Verfahren zur Klärung vielfältiger diagnostischer Fragestellungen an und organisiert Informationen zu Testpersonen und deren Ergebnisse in einer Datenbank. Die integrierte computerisierte Ergebnisauswertung und Darstellung der Testleistung ermöglicht eine fehlerfreie Verrechnung und sorgt so für Auswertungsobjektivität. Das WTS ermöglicht das Erstellen eigener beliebig variierbarer Testbatterien, die ohne Unterbrechung mit einer einheitlichen Benutzeroberfläche vorgegeben werden können. Von dieser Funktion wurde für die vorliegende Untersuchung Gebrauch gemacht, um eine standardisierte und störungsfreie Testung zu ermöglichen. Es kann also sowohl für die verwendeten Verfahren des WTS als auch für den ILICA aufgrund der computerisierten Vorgabe, von der Erfüllung des Gütekriteriums Objektivität ausgegangen werden. Die Konzeption als normierte Computertests gewährleistet weitgehende Testleiterunabhängigkeit sowie völlige Verrechnungssicherheit und Interpretationseindeutigkeit (Ortner, Kubinger, Schrott, Radinger & Litzenberger, 2011).

Die in dieser Studie vorgegebene Testbatterie besteht wie in Experiment 1 und 2 aus einem Leistungs- und einem Persönlichkeitsteil. Wie bereits in Abschnitt II.6. umfassend dargestellt, stellen Persönlichkeitsfragebögen aufgrund der hohen Verfälschbarkeit insbesondere im Kontext der Personalauswahl kein reliables und valides Mittel dar, um verlässliche und glaubwürdige Informationen von und über Bewerber zu erhalten. Im Persönlichkeitsteil dieser Testbatterie konnten die gleichen OPTs wie in den Vorgängerstudien verwendet werden. Die verwendeten Verfahren werden im Folgenden beschrieben. Zunächst werden die in dieser Studie verwendeten kognitiven Leistungstests vorgestellt, da sich hier die Testauswahl sowohl experimentell vorgesehen, als auch aus praktischen Gründen heraus, von den Vorgängerstudien unterscheidet.

2.2.1. AMT

Der adaptive Matrizentest (AMT) ist ein computergestütztes Verfahren zur sprachfreien Erfassung der allgemeinen Intelligenz. Anhand der mit diesem Verfahren erhobenen Fähigkeit zum logisch schlussfolgernden Denken wird ein Gesamtkennwert ausgegeben, der Aufschluss über die kognitive Leistungsfähigkeit geben soll. Das Itemmaterial entspricht dem Aufbau klassischer Matrizentests, die Testperson muss mehrere ähnlich gestaltete Aufgaben bearbeiten. Ein Item besteht dabei aus der Darstellung einer quadratischen Tafel, die sich aus neun Feldern zusammensetzt und darunter angeordneter Antwortmöglichkeiten. Acht Felder der Tafel enthalten abstrakte Symbole in Form geometrischer Konfigurationen. Das untere rechte Feld ist dabei stets frei gelassen und muss mit einem von den acht zur Verfügung stehenden Antwortmöglichkeiten ausgefüllt werden. Das Lösungssymbol soll die logischen Relationen und die Struktur der anderen acht Felder regelgeleitet komplettieren. Die Testvorgabe ist adaptiv und je nach Testverlauf werden aus dem 298 Items umfassenden Itempool etwa 20 bis 30 Items unterschiedlicher Schwierigkeit vorgegeben. Die Itemauswahl wird dabei laufend an die Leistungsfähigkeit oder, genauer gesagt, an die Schätzung des Fähigkeitsparameters der Testperson angepasst und ermöglicht so bei hoher Testökonomie einen maximalen Informationsgewinn. Den Items des AMTs liegt ein Konstruktionsrational zugrunde, welches explizit kognitionspsychologisch abgeleitet wurde, um die bei der Lösung von Matrizenaufgaben erforderlichen kognitiven Operationen im Itemmaterial korrekt

abzubilden und erfassen zu können. Die zugrundeliegenden Regeln des Konstruktionsrationalen müssen gefunden, abstrahiert und auf die vorgegebene Auswahl möglicher Antwortmöglichkeiten angewendet werden, um ein Item korrekt zu lösen.

Der AMT wurde in dieser Studie anstelle der in Experiment 2 vorgegebenen SPM verwendet. Bei den SPM handelt es sich ebenfalls um einen Matrizentest mit sehr ähnlichem Itemmaterial, allerdings erfolgt die Vorgabe im Gegensatz zum AMT nicht adaptiv, sondern klassisch mit steigenden Itemschwierigkeiten im Verlauf der Testvorgabe. Darüber hinaus differenziert der AMT besser im oberen Leistungsbereich, während der SPM besser im unteren Leistungsbereich differenziert und bei Personen mit hoher Leistungsfähigkeit Deckeneffekte zeigt.

Da adaptives Testen die Gültigkeit des Raschmodells voraussetzt, ist für den AMT die Reliabilität im Sinne der inneren Konsistenz als gegeben anzunehmen. Die Reliabilität lässt sich über die Angabe eines Zielstandardmessfehlers im WTS beliebig steuern und ist dann für alle Teilnehmer identisch. Die hier vorgegebene Standardform beendet die Testung, sobald der Messfehler 0,44 erreicht oder unterschritten hat. Bei Erreichen dieses kritischen Wertes entspricht die Reliabilität $r = 0,83$ (Cronbachs α). Das den Items des AMT zugrundeliegende Konstruktionsrational korreliert nach Hornke, Etzel & Küppers (2000) zu $r = 0,72$ mit den Schwierigkeitsparametern und belegt somit die Konstruktvalidität des Verfahrens. Die Durchführungsdauer des AMT liegt laut Manual bei etwa 54 Minuten. Die Bearbeitungszeit betrug in dieser Studie im Durchschnitt 39 Minuten.

2.2.2. INSBAT

Aus organisatorischen und finanziellen Gründen musste der in den Vorgängerstudien verwendete IST 2000-R in der hier vorliegenden Studie durch die Intelligenzstrukturbatterie (INSBAT) ersetzt werden.

In dieser Studie stand ausschließlich das WTS zur Verfügung, in dem der zum Hogrefe Testsystem gehörende IST 2000-R leider nicht enthalten ist. Angesichts der den finanziellen Rahmen überschreitenden Lizenzgebühren, die für Tests aus dem Hogrefe Testsystem (IST 2000-R) fällig geworden wären, und der demgegenüber problem- und kostenlosen Verfügbarkeit der INSBAT wurde eben diese anstelle des

IST 2000-R verwendet. Die INSBAT zählt standardmäßig zum Verfahrensinventar der ÖBB und es wurden dort stets gute Erfahrungen bei der Anwendung dieses Tests gemacht. Die Verwendung dieser Intelligenztestbatterie war dementsprechend erprobt und bewährt und aus Sicht der Verantwortlichen ebenfalls wünschenswert. Im Interesse einer unterbrechungsfreien und einheitlichen Vorgabe aller Verfahren mit Ausnahme des ILICAs innerhalb des WTS erwies es sich ebenfalls als günstig, die INSBAT zu verwenden. Ein weiterer Vorteil der INSBAT liegt darin begründet, dass aufgrund der adaptiven Testkonstruktion und der daraus resultierenden Testzeitverkürzung eine ökonomischere Testvorgabe bei gleichbleibender Ergebnisqualität erreicht wird. Die Beurteilung der Implementierung dieser Verfahrenstypen in die Testauswahl wurde bereits im Abschnitt II.5. besprochen und wird im Diskussionsteil der vorliegenden Arbeit erneut aufgegriffen und hinsichtlich der Vergleichbarkeit mit den Vorstudien und den möglichen Auswirkungen auf die Leistungen der hier herangezogenen Stichprobe umfassend und kritisch diskutiert. In der folgenden Darstellung der INSBAT wird stets vergleichender Bezug zum IST 2000-R genommen.

Die Intelligenzstrukturbatterie (INSBAT) stellt eine umfangreiche Sammlung von insgesamt 16 Einzelverfahren dar, die zu einer breit gefächerten allgemeinen Intelligenzdiagnostik verwendet werden können. Neben einem globalen Intelligenzniveau werden verschiedene Dimensionen der Intelligenzstruktur erhoben, die die allgemeine kognitive Leistungsfähigkeit erfassen und wiedergeben. Die Auswahl der zu erfassenden Kennwerte und Dimensionen ist flexibel, eine thematisch beliebige Kombination der einzelnen Verfahren ist vorgesehen. Die INSBAT ist für Testpersonen ab 14 Jahren vorgesehen und wird dezidiert für die Anwendung in der Personalpsychologie empfohlen. Sie stellt umfangreiche alters-, geschlechts- und bildungsspezifische Normtabellen zur Verfügung und erlaubt aufgrund des modularen Aufbaus eine Anpassung an nahezu jede geläufige Fragestellung der psychologischen Diagnostik. Diese kann beispielsweise eine allgemeine Intelligenzabklärung oder lediglich eine Begutachtung in spezifischen Bereichen der Intelligenzstruktur erfordern. Es stehen verschiedene vorgefertigte Testformen zur Verfügung, jedoch wurde hier aufgrund des Anforderungsprofils eine anwendungsspezifische variable Subtestzusammenstellung gewählt, die darüber hinaus speziell an den in Experiment 1 und 2 vorgegebenen IST 2000-R angelehnt wurde. Auf diese Weise wurden eine

größtmögliche Deckung der Kennwerte und eine hohe Übereinstimmung der mit beiden Verfahren erfassten Dimensionen erreicht. Die meisten Subtests des IST 2000-R finden dabei ein konzeptionelles Äquivalent in einem oder mehreren Subtests der INSBAT. *Tabelle 2* gibt Auskunft über die in Experiment 1 und 2 verwendeten Subtests des IST 2000-R und stellt diesen jene Subtests der INSBAT gegenüber, die mit dem Anspruch größtmöglicher theoretischer und inhaltlicher Entsprechung in der vorliegenden Studie verwendet wurden.

Tabelle 2:

Gegenüberstellung der Subtestauswahl des IST 2000-R und der INSBAT in Experiment 1, 2 und 3

Experiment 1	Experiment 2	Experiment 3
<i>IST 2000-R</i>	<i>IST 2000-R</i>	<i>INSBAT</i>
Verbal		
Satzergänzungen	Satzergänzungen	Wortbedeutungen
Analogien	Analogien	Verbal-deduktives Denken
Gemeinsamkeiten	Gemeinsamkeiten	Verbal-deduktives Denken
Numerisch		
Rechenaufgabe	Rechenaufgaben	Arithmetische Kompetenz
Rechenzeichen	Rechenzeichen	Arithmetische Flexibilität
Zahlenreihen	Zahlenreihen	Numerisch-induktives Denken
Gedächtnis		
	Merkfähigkeit figural	Visuelles Kurzzeitgedächtnis
	Merkfähigkeit verbal	Verbales Kurzzeitgedächtnis

Im Folgenden werden die sieben in dieser Studie verwendeten Subtests der INSBAT besprochen und dabei die Vergleichbarkeit zu den Subtests des IST 2000-R dargestellt. Es konnte dabei eine hohe Übereinstimmung der erfassten Variablen erreicht werden. Die Auflistung erfolgt in derselben Reihenfolge, in der die Subtests in der in beiden Versuchsgruppen verwendeten Testform vorgegeben wurden.

Numerisch-induktives Denken

Der Subtest *Numerisch-induktives Denken* zählt zum Aufgabenbereich logisch schlussfolgernden Denkens und entspricht in der theoretischen Konzeption und auch in der inhaltlichen Aufbereitung dem Subtest Zahlenreihen des IST 2000-R. Die Aufgabe besteht darin, den fehlenden Wert einer Zahlenreihe zu ergänzen und die Folge so anhand der zugrundeliegenden Regeln korrekt zu vervollständigen. Die Schwierigkeit besteht darin, diese Regeln zu erkennen und zu abstrahieren. Die Bearbeitung beansprucht die Kapazität des Arbeitsgedächtnisses.

Arithmetische Kompetenz

Der Subtest *Arithmetische Kompetenz* besteht aus Kopfrechenaufgaben unterschiedlichen Schwierigkeitsniveaus. Die so ermittelte Rechenfertigkeit zählt zum Bereich des Quantitativen Denkens, zur Bearbeitung ist der Umgang mit den vier Grundrechenarten erforderlich. Dieser Subtest enthält eine zusätzliche Speedkomponente, da die Aufgaben innerhalb eines Zeitfensters von 45 Sekunden gelöst werden müssen. Dieser Subtest findet in den Rechenaufgaben des IST 2000-R seine direkte Entsprechung. Er stellt Anforderungen an das deklarative Gedächtnis der Testpersonen sowie das Wissen um mögliche Rechenstrategien und deren Anwendung im prozeduralen Gedächtnis.

Verbal-deduktives Denken

Der Subtest *Verbal-deduktives Denken* erfasst die Fähigkeit zum sprachgebundenen deduktiven Denken. Aufgabe ist es, aus vorgegebenen Prämissen eine richtige Konklusion zu ziehen. Anhand zweier in einer logischen Beziehung zueinander stehenden Aussagen muss die zugrundeliegende Argumentation verstanden und auf vorgegebene Antwortmöglichkeiten angewendet werden. Um die Aufgabenstellung zu lösen, muss aus den Vorgaben eine logische Schlussfolgerung gezogen werden, die auch ungeachtet des Inhalts formal korrekt ist. Das Aufgabenmaterial besteht aus einem Katalog logischer Argumente, sogenannter Syllogismen. Dieser Aufgabentyp ist aus der philosophischen Logiklehre entlehnt und erfasst einen Teilaspekt des logisch schlussfolgernden Denkens.

Aus zweierlei Gründen wurde dieser Subtest vorgegeben. Zunächst war es wichtig, diesen Teilaspekt der fluiden Intelligenz im Sinne des Anforderungsprofils für den Auswahlprozess zu erfassen. Zusätzlich sollte eine Entsprechung zu den in

Experiment 1 und 2 verwendeten Subtests *Analogien* und *Gemeinsamkeiten* gefunden werden, um die Vergleichbarkeit zu den Vorläuferstudien zu gewährleisten. Es gestaltet sich trotz der hohen Ähnlichkeit in der theoretischen Konzeption des IST 2000-R und der INSBAT schwierig, die Äquivalenz dieser drei Subtests auf dieser Ebene zu begründen. Die hierarchisch übergeordneten Dimensionen aller drei Subtests in ihren jeweiligen Verfahren weisen jedoch eine hohe Deckung auf, es handelt sich in beiden Testbatterien jeweils um sprachgebundenes logisch schlussfolgerndes Denken. Diese Dimension wird in beiden Testbatterien lediglich auf sehr unterschiedliche Weise operationalisiert. Aufgrund der Sprachgebundenheit und der Subsummierung unter eine vergleichbare Dimension sowohl im IST 2000-R als auch in der INSBAT, kann für diese drei Subtests trotz großer inhaltlicher Unterschiede eine Verwandtschaft angenommen werden. Somit erscheint es gerechtfertigt, den Subtest *Verbal-deduktives Denken* als Äquivalent zu den Subtests *Analogien* und *Gemeinsamkeiten* des IST 2000-R zur Erfassung der Dimension des sprachgebundenen logisch schlussfolgernden Denkens zu verwenden.

Verbales Kurzzeitgedächtnis

Das verbale Kurzzeitgedächtnis wird mit dem gleichnamigen Subtest erfasst und entspricht konzeptionell dem Subtest Merkfähigkeit verbal des IST 2000-R. Die sprachgebundene Gedächtnisleistung wird hier mithilfe einer simulierten Busfahrt auf einem abstrakten Stadtplan abgeprüft. Die Aufgabe der Testperson besteht darin, sich die korrekte Abfolge der Haltestellen und deren mehr oder minder komplexe Namen zu merken und sie anschließend zu reproduzieren. Neben der Fähigkeit des seriellen Lernens ist hier vor allem eine effiziente kurzfristige Lernstrategie erforderlich und in besonderem Maße zielführend.

Arithmetische Flexibilität

Der zweite Subtest des Fähigkeitsbereich Quantitatives Denken ist die *Arithmetische Flexibilität*. Diese Dimension wird äquivalent zum Subtest Rechenzeichen des IST 2000-R durch die Vorgabe mathematischer Gleichungen abgeprüft. Die Testpersonen müssen die fehlenden Rechenzeichen der vier Grundrechenarten ergänzen, um die Gleichung korrekt lösen zu können. Zur Lösung erforderlich ist ein gutes Verständnis für grundlegende mathematische Rechenoperationen und die Fähigkeit diese Prinzipien auf lösungsrelevante Informationen aus dem Langzeitgedächtnis anzuwenden. Dieser Aufgabentyp stellt

hohe Anforderungen an die kognitive Flexibilität und die Kapazität des Arbeitsgedächtnisses und wurde im Laufe der Testung von vielen Teilnehmern als der schwierigste und am meisten belastende Subtest beschrieben.

Visuelles Kurzzeitgedächtnis

Das visuelle Kurzzeitgedächtnis wird ähnlich zum verbalen Kurzzeitgedächtnis anhand eines abstrakten Stadtplans simuliert. Diesmal besteht die Aufgabe der Testperson darin, die Position kurzzeitig dargebotener Symbole zu behalten und anschließend nach deren Verschwinden auf dem leeren Stadtplan zu reproduzieren. Die Symbole, deren Bedeutung vorab erklärt wurde, repräsentieren dabei etwa einen Bahnhof oder eine Bücherei. Die so erfasste Dimension entspricht der Merkfähigkeit figural des IST 2000-R.

Wortbedeutungen

Der Subtest *Wortbedeutungen* erfasst das Wortverständnis und erlaubt Rückschlüsse auf den Wortschatz einer Testpersonen. Dabei wird ein Wort vorgegeben, und die Testperson muss aus vier verschiedenen Antwortmöglichkeiten jenes Wort auswählen, welches dem vorgegebenen inhaltlich am nächsten kommt, also eine möglichst genaue Entsprechung darstellt. Eine vergleichbare Dimension wird im IST 2000-R mit dem Subtest Satzergänzungen erfasst.

Es wird deutlich, dass die Übereinstimmung der Subtests des IST 2000-R und der INSBAT in vertretbarem Umfang gewährleistet ist, letztendlich basieren beide Verfahren auf einem hierarchischen Generalfaktormodell, dessen Intelligenzkonzept in vielen Punkten hohe Ähnlichkeit aufweist. Somit ist ein Vergleich der mit den beiden Verfahren erhobenen Kennwerte und Dimensionen zulässig. Den wohl entscheidenden Unterschied zwischen beiden Verfahren stellt die adaptive Testkonstruktion der INSBAT im Gegensatz zur klassischen Vorgabe mit ansteigenden Itemschwierigkeiten im Testverlauf im IST 2000-R dar.

IST 2000-R und INSBAT befinden sich hinsichtlich der Gütekriterien auf ähnlich hohem Niveau und sorgen auch in dieser Hinsicht für Vergleichbarkeit der Studien. Sämtliche hier verwendeten Subtests der INSBAT werden adaptiv vorgegeben. Es ist also die Reliabilität im Sinne der internen Konsistenz aufgrund der Gültigkeit des Rasch-Modells als gegeben anzunehmen. Wie auch beim AMT kann die Reliabilität

frei durch die Angabe eines kritischen Messfehlerwertes im WTS adjustiert werden. In der hier verwendeten Form variieren die vorgegebenen Zielreliabilitäten, bei deren Erreichung die Testung im Sinne eines optimalen Verhältnisses von Testökonomie zu Messgenauigkeit beendet wird, zwischen $r = 0,73$ und $r = 0,75$ (Cronbachs α). Laut Manual beziffern Studien die Korrelationen der adaptiven Subtests zugrundeliegenden Konstruktionsrationalen mit den Itemparametern auf Werte zwischen $r = 0,70$ und $r = 0,95$, was als Beleg für die Konstruktvalidität ausgelegt wird. Die Durchführungsdauer der INSBAT ist je nach Subtestauswahl und Voreinstellungen verschieden. In der hier vorgegeben Version lag diese laut WTS bei etwa 99 Minuten. Die mittlere Bearbeitungszeit betrug in dieser Studie jedoch 115 Minuten.

Im Folgenden werden jene drei Verfahren vorgestellt, die den Persönlichkeitsteil der vorgegebenen Testbatterie repräsentieren. Bei den verwendeten Verfahren handelt es sich ausschließlich um Objektive Persönlichkeitstests (OPT) sensu Cattell (1958). Wie bereits im Theorieteil ausgeführt, können diese Verfahren verwendet werden, um Persönlichkeitscharakteristika möglichst objektiv zu messen. Dabei wird Verfälschungstendenzen, die bei Persönlichkeitsfragebögen evident und besonders bei der Eignungsdiagnostik in Personalauswahlprozessen problematisch sind, effektiv entgegengewirkt und so beispielsweise der Einfluss sozial erwünschten Antwortverhaltens reduziert. Mithilfe dieser meist computerisierten Verfahren wird das beobachtbare Verhalten während der Bearbeitung erfasst. Dabei verschleiern die vorgegebenen Tests die eigentliche Messintention durch die Art des Aufgabenmaterials.

2.2.3. AHA

Die Arbeitshaltungen sind ein aus drei Subtests bestehender OPT, der verschiedene Aspekte des Arbeitsverhaltens bei Leistungsanforderungen erfasst. Erhoben werden Impulsivität und Reflexivität des Arbeitsstils, das Anspruchs- und Leistungsniveau sowie die Frustrationstoleranz und Leistungsmotivation. Die AHA werden von den Autoren explizit zur Verwendung im Personalwesen und bei der Berufsberatung empfohlen.

Flächengrößen Vergleichen

Im ersten Subtesttest *Flächengrößen Vergleichen* werden der Testperson zwei skizzierte Flächen gleichzeitig dargeboten. Diese Flächen müssen innerhalb eines Zeitfensters von 30 Sekunden hinsichtlich ihrer Größenordnung eingeschätzt werden. Die Testperson ist aufgefordert, jene Fläche anzuklicken, die als größer wahrgenommen wird. Erfasst wird dabei zunächst die Exaktheit bei der Bearbeitung, also der Anteil korrekt gelöster Aufgaben an der Gesamtzahl der Aufgaben. Ein Maß für die Entschlussfreudigkeit der Testperson wird anhand der Summe korrekter und falscher Antworten errechnet. Das Verhältnis korrekter, unkorrekter und nicht beantworteter Aufgaben wird als Indikator für die Impulsivität beziehungsweise Reflexivität des Arbeitsstils gewertet.

Symbole Kodieren

Im darauf folgenden Subtest *Symbole Kodieren* muss die Testperson in mehreren zeitlich limitierten Durchgängen anhand eines Kodierschlüssels fortlaufend vorgegebene abstrakte Symbole korrekt zuordnen. Die Testperson wird dann vor jedem weiteren Durchgang aufgefordert anzugeben, wie viele Symbole sie ihrer Einschätzung nach im folgenden Durchgang korrekt zuordnen wird. Auf diese Weise wird, neben der Erfassung des Leistungsniveaus anhand der Summe korrekt zugeordneter Symbole, das Anspruchsniveau und die Frustrationstoleranz über die Differenz zwischen angegebener und tatsächlich gelöster Aufgabenmenge über den Testverlauf erfasst.

Figuren Unterscheiden

Im letzten Subtest *Figuren Unterscheiden* müssen die Testpersonen eine repetitive und monotone Kodieraufgabe bearbeiten. Diese besteht darin, in fortlaufenden Reihen aus jeweils vier gleichen und einem abweichenden Symbol die unregelmäßigen Symbole auszuwählen. Anhand des Durchhaltevermögens bei dieser wenig anspruchsvollen Aufgabe wird die Leistungsmotivation der Testperson erfasst.

Im Rahmen der vorliegenden Studie wurde, wie auch in den Experimenten 1 und 2, eine von der Originalversion abweichende Variante dieses letzten Subtests vorgegeben. Mit dem Ziel einer verkürzten Gesamttestdauer wurde dieser Subtest mit einer verringerten Bearbeitungszeit von maximal 15 Minuten vorgegeben. Bei Erreichen dieser Zeitmarke wurde der Test automatisch abgebrochen. In der Originalversion wird dieses Kriterium erst nach 30 Minuten erreicht.

Wie bereits im Abschnitt II.6. angesprochen, gestaltet sich die Bestimmung der Reliabilität und die Validierung von OPTs als schwierig. Angaben zur Reliabilität der AHA werden von den Autoren mit der Begründung einer nicht sinnvollen Bestimmbarkeit dieser Kennwerte abgelehnt. Studien zur Validität (Kubinger & Hofmann, 1998; Kubinger & Ebenhöf, 2007) kamen auch in Ermangelung geeigneter konstruktnaher Verfahren zu widersprüchlichen und wenig aussagekräftigen Ergebnissen. Die Konstruktvalidität wird dennoch durch die theoretische Fundierung der Kennwerte und eine faktorenanalytische Überprüfung der Faktorenstruktur als gegeben angenommen. Die durch den verkürzten Subtest *Figuren Unterscheiden* leicht veränderte Version der AHA hat, adaptiert aus den Informationen des Manuals, eine Durchführungsdauer von etwa 30 Minuten. Die Bearbeitungszeit betrug in dieser Studie durchschnittlich 19 Minuten.

2.2.4. BACO

Der computergestützte objektive Persönlichkeitstest BACO wird verwendet, um die Belastbarkeit einer Person anhand ihres Verhaltens bei der Bearbeitung simulierter, meist arbeitsbezogener Tätigkeiten zu erfassen. Die in dieser Studie vorgegebenen Subtests erfassen die Belastbarkeit bei Aufgabenkollision und die Belastbarkeit bei Verhinderung des planmäßigen Vorgehens. Die Bearbeitung dieser Subtests gibt Aufschluss über unterschiedliche Komponenten der Belastbarkeit einer Testperson, sie erfassen das Arbeits- und Prozessverhalten in simulierten Belastungssituationen. Anhand des Verhaltens in diesen Situationen werden Rückschlüsse auf die Leistungsfähigkeit der Testpersonen unter belastenden Umständen gezogen.

Aufgabenkollision

Der Subtest *Aufgabenkollision* ist eine computerisierte Umsetzung von Postkorbszenarien wie sie in Assessment-Centern bei der Personalauswahl Verwendung finden. Die Testperson muss eine als Hauptaufgabe deklarierte Tätigkeit ausführen, deren vollständige Bearbeitung fortlaufend durch diverse Nebenaufgaben unterbrochen wird. Die effektive und vollständige Bearbeitung der Hauptaufgabe wird

so verhindert. Anhand mehrerer Kennwerte wird erfasst, wie sich die Testperson in dieser Belastungssituation verhält. Erhoben werden die Quantität, die Qualität und die Effizienz bei der Bearbeitung der Hauptaufgabe sowie der Nebenaufgaben. Zusätzlich wird die Ausgewogenheit im Verhältnis der Bearbeitung beider Tätigkeiten und das Durchhaltevermögen erfasst, welches über die Konstanz in der Leistung über die gesamte Subtestdauer definiert ist.

Verhinderung des planmäßigen Vorgehens

Die Belastungssituation bei *Verhinderung des planmäßigen Vorgehens* wird mit Hilfe eines Labyrinthtests simuliert. Die Testperson muss in einem anscheinend statischen und mit vielen offensichtlichen Lösungswegen ausgestatteten Labyrinth eine Spielfigur Richtung Ausgang dirigieren. Das Labyrinth bietet zwar zu Beginn viele offene Wege zum Ziel, diese werden jedoch - sukzessive und dynamisch an das Verhalten der Testperson angepasst - durch plötzlich auftauchende Labyrinthelemente blockiert. So wird eine Zielerreichung auf dem gerade eingeschlagenen Weg verhindert. Zeigt die Testperson ein hohes Maß an Persistenz bei der Bearbeitung, wird nach 42 blockierten Lösungswegen die weitere Bearbeitung nicht mehr behindert und ein Erreichen des Ziels ermöglicht. Erfasst werden neben dem Grad der Zielerreichung, die Geschwindigkeit und das Verhalten bei der Exploration möglicher Lösungswege. Verharrt die Testperson angesichts auftretender Blockaden einige Zeit oder beschreitet sie öfters bereits eingeschlagene und blockierte Wege, so wird dies als zögerliches und verunsichertes Verhalten interpretiert.

Im Gegensatz zu dem in Experiment 1 und 2 verwendeten BACO-D wurde der BACO in der vorliegenden Studie in einer überarbeiteten und neu gestalteten Version vorgegeben. Der BACO ist seit 2011 in das WTS integriert und erfuhr in Folge dessen eine grafische und programmiertechnische Überarbeitung. Das Verfahren, die Messintention und das Aufgabenmaterial selbst wurden dabei nicht verändert, lediglich die grafische Oberfläche wurde an das Layout des WTS angepasst und es wurden einige Programmierfehler behoben. Die Vorgabe des BACOs erfolgte im Gegensatz zu den vorangegangenen Experimenten 1 und 2 nahtlos anschließend an die anderen Tests der vorgegebenen Testbatterie und ohne Unterbrechung innerhalb der Vorgabemaske des WTS. So wurde ein unterbrechungs- und störungsfreier Testverlauf garantiert. Anhang IV.4. zeigt beispielhaft die Unterschiede in der

grafischen Aufbereitung des in Experiment 1 und 2 verwendeten BAcO-D und des BACO in seiner derzeitigen, in der vorliegenden Studie verwendeten Form.

Mit Ausnahme des Testkennwerts Geschwindigkeit aus dem Subtest Belastbarkeit bei *Aufgabenkollision* sind die beiden Testversionen BAcO-D und BACO auch hinsichtlich der Auswertung identisch. Die Auswirkungen dieser Abweichung wurden bei den statistischen Analysen berücksichtigt. In Abschnitt III.3.2. wird dieser Umstand erläutert und bei der anschließenden Interpretation der Ergebnisse berücksichtigt.

Die ausgewählten Subtests, insbesondere die Belastbarkeit bei *Verhinderung des planmäßigen Vorgehens*, erweisen sich nach Einschätzung der Autoren als ungeeignet für eine verlässliche Bestimmung der Reliabilität. Die *split-half*-Reliabilität der Kennwerte Quantität Hauptaufgabe und Effizienz Nebenaufgabe des Subtests Belastbarkeit bei *Aufgabenkollision* wird dennoch mit $r = 0,57$ und $r = 0,91$ angegeben (Spearman & Brown). Zur Validität des Verfahrens wird zunächst berichtet, dass die mit dem BACO erfasste Belastbarkeit zwar nicht mit Persönlichkeitsfragebögen erfasst werden kann (Kieweg, 2004), sich aber eindeutig durch die Testung psychophysiologischer Belastungsreaktionen nachweisen lässt (Ortner, 2001). Für den Kennwert Effizienz der Nebenaufgabe und sämtliche Kennwerte des Subtests Belastbarkeit bei *Verhinderung des Planmäßigen Vorgehens* wird die inhaltliche Gültigkeit, konzeptionell bedingt, als gegeben angenommen. Die Bearbeitungszeit des BACOs richtete sich nach den ausgewählten Subtests. Die beiden hier vorgegebenen Subtests haben laut Manual eine Gesamtdauer von etwa 21 Minuten. Die mittlere Durchführungsdauer betrug in dieser Studie 17 Minuten.

2.2.5. ILICA

Der ILICA wurde als letzter von drei OPTs im Persönlichkeitsteil der Testbatterie vorgegeben. Beim ILICA handelt es sich um eine Computersimulation zur Erfassung der Selbstorganisationsfähigkeit und des strategischen Planungsverhaltens. Aufgabe der Testperson ist es, einen fiktiven Tagesplan zu gestalten, der den letzten Tag vor einer Urlaubsreise strukturieren soll. Dafür werden der Testperson verschiedene Aufgaben fest vorgegeben, die es zu erledigen gilt, und darüber hinaus muss mithilfe einer vielfältigen Auswahl an optionalen Tätigkeiten der Tagesplan aufgefüllt werden. Erfasst werden daraufhin der Grad der Erfüllung der vom Verfahren vorgegebenen und

der selbst vorgenommenen Aufgaben. Die Art der verwendeten Hilfsmittel und das Ausmaß ihrer Verwendung werden ebenfalls erhoben. Unter Einbezug aller erhobenen Variablen erfolgt letztendlich eine Typisierung der Testperson anhand der drei Eigenschaften Flexibilität, Ablenkbarkeit und Zielorientierung.

Eine Bestimmung der Reliabilität im Sinne der internen Konsistenz oder einer re-test-Reliabilität ist laut Manual weder möglich noch zielführend. Während sich die diskriminante Validität anhand geringer Korrelation mit diversen kognitiven Leistungstests noch recht gut belegen lässt, kann zur konvergenten Validität in Ermangelung konstruktnaher Verfahren keine zufriedenstellende Aussage gemacht werden. Die Kennwerte des ILICAs korrelieren untereinander lediglich gering, und das Verfahren konnte bei der Anwendung gute Ergebnisse hinsichtlich der prognostischen Validität erreichen. Die Testdauer des ILICAs wird im Manual mit 40 bis 50 Minuten beziffert. Die Bearbeitungszeit wird vom ILICA nicht eigens erfasst und kann daher hier nicht angegeben werden. Erfahrungsgemäß dürfte die Bearbeitungszeit des ILICAs in dieser Studie jedoch wie bei den anderen Verfahren mit Ausnahme der INSBAT ebenfalls tendenziell kürzer als im Manual vorausgesagt gewesen sein. Realistisch dürfte von etwa 20 bis 30 Minuten ausgegangen werden.

2.3. Durchführung der Untersuchung

Die Testbatterie wurde den zwei Versuchsgruppen zu je 35 Testpersonen mit jeweils unterschiedlichen Testreihenfolgen vorgegeben. Die Testungen waren Teil des Bewerberauswahlprozesses eines ÖBB Trainee-Programms für angehende Führungskräfte der Bereiche Logistik, Controlling, Bauingenieurswesen, Betriebswirtschaft und Human Resources.

Der gesamte Auswahlprozess zog sich über mehrere Monate hin, die Testungen wurden dabei innerhalb von 6 Wochen durchgeführt. Die Testungen erfolgten direkt im Anschluss an Bewerbungsgespräche, die von Mitarbeitern der ÖBB durchgeführt wurden. Die Verfahren waren auf den verwendeten Computern vorbereitet und liefen weitgehend automatisiert innerhalb des WTS ab. Im Vorfeld wurden die Systeme mehrfach getestet und der korrekte Ablauf der Testung überprüft. Die unterschiedlichen Testreihenfolgen wurden vorprogrammiert. Lediglich die Vorgabe des ILICAs bedurfte einer manuellen Aktivierung durch die Testleiter. Die Teilnehmer

wurden randomisiert den jeweiligen Untersuchungsbedingungen zugeteilt. Dies wurde durch eine zufällige Zuweisung der Testpersonen zu den jeweiligen Testplätzen realisiert, an denen jeweils lediglich eine der beiden Testreihenfolgen aktiviert war. Von den zehn Testplätzen waren jeweils fünf mit derselben Testreihenfolge programmiert. Ein zusätzlicher elfter Testplatz war variabel mit beiden Testreihenfolgen ausgestattet.

Innerhalb von 6 Wochen in den Monaten Mai, Juni und Juli 2012 wurden sämtliche Testungen im Hauptgebäude der ÖBB Holding in Wien durchgeführt. Ein speziell für Computertestungen konzipierter und ausgerüsteter Testraum stand für die Testungen zur Verfügung. Der Raum war klimatisiert, gut beleuchtet und bot optimale Bedingungen für eine störungsfreie und ruhige Testdurchführung. Um die Vorgabe in größtmöglichem Maße zu standardisieren und kontrollieren zu können, wurden alle Personen in derselben Umgebung getestet. Der Testraum verfügte über insgesamt 13 Einzelplätze, von denen 11 für diese Studie genutzt wurden. Die Testplätze waren soweit abgeschirmt, dass die Testpersonen weitestgehend von den anderen Paralleltestungen isoliert und unbeeinflusst waren. Es bestand zusätzlich die Möglichkeit, zur Erhöhung der Geräuschisolation Kopfhörer anzulegen, diese Option wurde jedoch von keiner Testperson wahrgenommen, wodurch auch hier einheitliche Bedingungen gegeben waren. Die Testungen selbst wurden an insgesamt elf Laptops durchgeführt, die vom Arbeitsbereich Psychologische Diagnostik des Instituts für Psychologie der Universität Wien und der Schuhfried GmbH gemeinsam zu Studienzwecken zur Verfügung gestellt wurden. Die Laptops waren mit vorinstalliertem WTS ausgerüstet und konnten problemlos mit der in den Testräumen vorhandenen Peripheriegeräten verbunden werden. Alle Teilnehmer hatten somit die gleichen Monitore bei identischer Auflösung sowie die gleichen Computermäuse und Tastaturen, womit auch in dieser Hinsicht eine größtmögliche Standardisierung garantiert wurde.

Der zeitliche Ablauf des Auswahlprozesses war für alle Teilnehmer gleich. Über den Tag verteilt wurden zwischen 8 und 17 Uhr mit maximal zwei Bewerbern unabhängig voneinander parallele Bewerbungsgespräche von Mitarbeitern der Personalabteilung der ÖBB geführt. Diese Interviews dauerten etwa 45 bis 60 Minuten. Direkt im Anschluss wurden die Teilnehmer in den in der Nähe befindlichen Testraum geführt und dort von zwei Testleitern in Empfang genommen, die ihnen einen Testplatz zuwiesen. Es wurde darauf geachtet, dass stets Testleiter beiderlei Geschlechts anwesend waren, um etwaige geschlechtsspezifische Testleitereffekte

auszuschließen. Die Testung wurde dann nach Aufnahme der biographischen und demographischen Daten sowie einer kurzen und standardisierten mündlichen Instruktion gestartet. Die Testpersonen wurden darüber informiert, dass die folgenden Tests einen immanenten Teil des Auswahlprozesses darstellen, dass sie über theoretisch unbegrenzte Arbeitszeit verfügen, diese jedoch nicht notwendigerweise auszuschöpfen ist. Ein zügiger und konzentrierter Arbeitsstil wurde empfohlen. Im Hinblick auf die erwartungsgemäß hohe Anforderung durch die adaptiven Testteile wurden die Teilnehmer dahingehend instruiert, sich von schwierigen Aufgaben, die sie eventuell gar nicht oder zumindest nicht in absehbarer Zeit lösen können, nicht verunsichern zu lassen. Es wurde darauf hingewiesen, dass darauf folgend auch wieder Aufgaben vorkommen würden, die für sie lösbar seien. Die Testpersonen wurden ebenfalls darüber informiert, dass es ihnen gestattet sei kurze Pausen von maximal fünf Minuten zu machen. Pausen seien indes jedoch lediglich zwischen und nicht während der Bearbeitung einzelner Verfahren möglich. Den Teilnehmern wurde Trinkwasser zur Verfügung gestellt und beständig von den Testleitern nachgefüllt. Den Teilnehmern wurde versichert, jederzeit Fragen stellen zu können, und sie wurden gebeten, sich bei Verständnisschwierigkeiten oder technischen Problemen an die dauerhaft anwesenden Testleiter zu wenden. Es wurde jedoch darauf hingewiesen, dass die Tests über ausführliche und leicht verständliche Instruktionen und teilweise auch Übungsaufgaben verfügten. Nachfragen waren im Gesamtverlauf eher selten, Verständnisschwierigkeiten tauchten kaum auf, auch technische Probleme blieben ausgesprochen selten.

Nach Abschluss der Testungen wurden die Teilnehmer aus dem Raum begleitet, draußen verabschiedet und allfällige Fragen von den Testleitern beantwortet. Die Teilnehmer durften daraufhin nach Hause gehen und hatten diesen Teil des Auswahlprozesses erledigt.

Die Testergebnisse wurden daraufhin am Abend nach Verabschiedung des letzten Teilnehmers ausgewertet, zu adäquaten Normstichproben ins Verhältnis gesetzt und den Verantwortlichen der ÖBB in aufbereiteter Form zur Verwendung im weiteren Auswahlprozess zur Verfügung gestellt. Es wurden umfassende Rückmeldebögen mit ausführlichen Erläuterungen der verwendeten Verfahren, der einzelnen Testkennwerte und deren Interpretation erstellt. Eine Rückmeldung der Testergebnisse an die Testpersonen erfolgte auf Anfrage in telefonischer Form. Eine exemplarische Darstellung der Rückmeldebögen ist dem Anhang IV.1. zu entnehmen.

2.4. Stichprobe

Die a-priori Berechnung eines optimalen Stichprobenumfangs zur Berechnung einer MANOVA wurde unter Annahme der Signifikanzniveaus von $\alpha = 0,05$ und $\beta = 0,20$ mit dem Programm G*Power in der Version 3.1.5 (Faul, Erdfelder, Lang & Buchner, 2009) durchgeführt. Hinsichtlich der a-priori anzugebenden Effektstärke wurde auf die signifikanten Ergebnisse aus Experiment 1 zurückgegriffen und von einer mindestens zu erwartenden Effektstärke von $\eta^2 = 0,450$ ausgegangen. Ein geringerer Effekt würde als nicht praktisch relevant zu bezeichnen sein. Es wurde somit ein gleichwertiger bis stärkerer Effekt vorausgesetzt, um als inhaltlich und statistisch bedeutsam anerkannt werden zu können. Für die Berechnungen wurde der auch von SPSS genutzte Schätzalgorithmus nach Muller und Peterson (1984) verwendet. Es wurde so ein optimaler Stichprobenumfang von $N = 70$ ermittelt. Mit einer Stichprobengröße von $N = 35 \times 2 = 70$ ist der in dieser Studie erhobene Stichprobenumfang somit als adäquat anzusehen.

Die Stichprobe setzt sich aus vorselektierten Bewerbern um Stellen in einem Trainee-Programm der ÖBB zusammen, die zur Teilnahme am Auswahlprozess eingeladen wurden. Voraussetzung für eine Stellenbewerbung war ein abgeschlossenes fachspezifisches Hochschulstudium. Erwartungsgemäß handelte es sich bei den Bewerbern größtenteils um die typische von der ÖBB angesprochene und erwünschte Zielgruppe eines Trainee-Programms. Dies sind vor allem junge Hochschulabgänger mit wenig Berufs- und Führungserfahrung, die das Trainee-Programm als Einstiegsmöglichkeit und Karrieresprungbrett in Unternehmen ansehen. Trainees gelten als potenzielle Mitarbeiter mit hoher Leistungsfähigkeit, hoher Qualifikation und hoher fachlicher Kompetenz, die im Laufe des Trainee-Programms auf kommende Führungspositionen vorbereitet werden sollen.

Das Geschlechterverhältnis war in der vorliegenden Studie im Unterschied zu Experiment 1 und insbesondere zu Experiment 2, in welchem die Stichprobe ausschließlich aus Männern bestand, nahezu ausgewogen. 39 Frauen und 31 Männer nahmen an den Testungen teil. Das Geschlecht wurde in dieser Studie bei den Analysen nicht als Faktor berücksichtigt, da ein Einfluss weder theoretisch begründet erwartet noch hypothesengeleitet überprüft werden sollte. Das Geschlechterverhältnis insgesamt und innerhalb der Gruppen stellt sich wie in *Tabelle 3* ersichtlich dar. Die

Verteilung der Geschlechter auf die Versuchsgruppen erfolgte aufgrund der Randomisierung ebenfalls zufällig. Gemessen am Geschlechterverhältnis in der Gesamtstichprobe ist die Verteilung innerhalb der Gruppen als akzeptabel anzusehen.

Tabelle 3:

Verteilung der Geschlechter pro Gruppe und insgesamt

	<i>Frauen</i>	<i>Männer</i>
Gruppe 1 LT → OPT	18	17
Gruppe 2 OPT → LT	21	14
Gesamt	39	31

Die Altersspanne der Teilnehmer reichte von 22 bis 32 Jahren. Der Altersdurchschnitt der Testpersonen lag bei 25,8 Jahren mit einer Standardabweichung von 2,158. Damit charakterisiert sich die Stichprobe im Gegensatz zu Experiment 1 und 2 als vergleichsweise jung. Am Auswahlprozess und der Testung nahmen kaum Personen höheren Alters teil, da sie auch als Zielgruppe nicht explizit angesprochen wurden.

Anhand der Altersverteilung ist auf eine weitere Charakteristik der Stichprobe zu schließen, die insbesondere im Vergleich mit den Stichproben aus Experiment 1 und 2 von essentieller Bedeutsamkeit ist. Wer verhältnismäßig wenige Jahre in einem Unternehmen tätig ist, verfügt in der Regel über weniger Führungserfahrung, da das Bekleiden einer Führungsposition neben der fachlichen Qualifikation auch vom Alter und der Dauer der Zugehörigkeit zum Unternehmen abhängig ist (Schneider & Stein, 2006). Aufgrund des geringen Alters und der zusätzlichen Voraussetzung, ein fachspezifisches Hochschulstudium absolviert zu haben, verfügte der Großteil der Teilnehmer nur über wenig und eingeschränkt einschlägige Berufserfahrung. Nur vereinzelt waren Teilnehmer im Vorfeld bereits mehrere Jahre in einem bildungs-, qualifikations- und anforderungsadäquaten Beruf tätig gewesen. Auch in den informellen Unterhaltungen im Anschluss an die Testungen gab ein Großteil der

Teilnehmer an, dass dies für sie die erste Teilnahme an einem Auswahlprozess für eine ausbildungsadäquate Stelle mit Vollzeitanstellung gewesen sei. Lediglich einige wenige Teilnehmer waren bereits in einer vergleichbaren Position tätig wie diejenige, auf die sie im Rahmen des angebotenen Trainee-Programms vorbereitet werden sollten. Diese Information ließ sich auch anhand der Bewerbungsunterlagen der Teilnehmer verifizieren.

Das Bildungsniveau der Teilnehmer war aufgrund der Bewerbungsmodalitäten und der Voraussetzung eines abgeschlossenen fachspezifischen Hochschulstudiums als vergleichsweise hoch anzunehmen. Die Teilnehmer verfügten mindestens über einen Bachelor-, in den meisten Fällen sogar über einen Master- oder Diplomabschluss in ihren jeweiligen Fachrichtungen. Somit charakterisiert sich die Stichprobe hinsichtlich ihres Bildungsniveaus als sehr homogen. Für die Rückmeldungen wurde, sofern verfügbar, auch stets eine alters- und bildungsadäquate Normstichprobe herangezogen. Das hohe Leistungsniveau zeigte sich bei der Auswertung oftmals anhand der im Vergleich zur herangezogenen Normstichprobe hohen bis sehr hohen Werte im Leistungsbereich. Das teils beeindruckende Leistungsniveau der Teilnehmer ist beispielsweise anhand des exemplarisch beigefügten Rückmeldebogens in Anhang IV.1. ersichtlich und stellt keinen Einzelfall dar.

Eine weitere für das Verständnis und die Interpretation der Ergebnisse relevante Eigenheit dieser Stichprobe hängt sowohl mit dem Alter als auch mit dem Bildungsniveau der Stichprobe zusammen. Es ist anzunehmen, dass die Teilnehmer aufgrund der zeitlichen Nähe zu ihrem Studienabschluss über ein hohes Maß an Routine und Erfahrung mit Leistungstests und Prüfungssituationen verfügen. Während Personen, die schon längere Zeit im Unternehmenskontext beschäftigt sind, kaum mit Prüfungssituationen, die in etwa mit den Testungen in dieser Studie vergleichbar sind, in Berührung kommen, haben die Teilnehmer hier erst kürzlich mehrere Jahre konstanter Prüfungssituationen im Rahmen ihres Studiums hinter sich. Bei der Interpretation der Ergebnisse wird dies zu berücksichtigen sein, da ein nicht unerheblicher Einfluss dieser Erfahrung auf die Leistung und das Verhalten der Teilnehmer in Testsituationen im Vergleich zu den Stichproben aus Experiment 1 und 2 zu erwarten ist.

Zusammengenommen zeigt die Darstellung der Stichprobe in diesem Abschnitt, dass sie sich, wie bereits im Abschnitt III.1. angedeutet, in relevanten Eigenschaften von der aus Experiment 2 unterscheidet. Absicht der vorliegenden Studie war es, eine vom Bildungsniveau zu Experiment 2 vergleichbare Stichprobe heranzuziehen, um dieses Experiment mit dem AMT anstelle der SPM replizieren zu können. Es konnte jedoch leider nur auf diese Stichprobe mit einem höheren Bildungsniveau zurückgegriffen werden. Dennoch gibt es bedeutsame Gemeinsamkeiten der beiden Stichproben, beispielsweise die geringe Führungserfahrung, auch wenn diese aus unterschiedlichen Umständen resultiert, und eine Tätigkeit in einem technisch orientierten Unternehmen, die für die Diskussion der Ergebnisse interessant sind und zumindest in gewissem Grade eine Vergleichbarkeit ermöglichen. Andererseits unterscheiden sich die Stichproben neben dem Bildungsniveau vor allem in der Altersspanne und der Tätigkeitsbereiche, die hier weniger handwerklich sind, aber auch in der Geschlechterverteilung, was eine eindeutige Einschränkung der Aussagekraft bedeutet.

3. Auswertung

Die deskriptiven und inferenzstatistischen Auswertungen der Daten wurden mithilfe der Software IBM SPSS Statistics in der Version 20 sowie R in der Version 2.15.2 (R Development Core Team, 2012) durchgeführt. Für sämtliche Berechnungen wurde, analog zu Experiment 1 und 2 und wie bei der Stichprobenvorausberechnung zur Untersuchungsplanung vorgesehen (Abschnitt III.2.4.), das Risiko für den Fehler 1. Art auf 5% festgelegt, als Signifikanzniveau somit $\alpha = 0,05$ gewählt.

Zur Klärung der durch die Hypothesen 1 und 2 formulierten Fragestellung wurden die erhobenen Kennwerte der Versuchsgruppen einer inferenzstatistischen Analyse unterzogen. Da die unterschiedlichen Testreihenfolgen durch die beiden Versuchsgruppen operationalisiert wurden, lässt sich anhand von Mittelwertsdifferenzen in den Variablen beider Versuchsgruppen eine Aussage über das Vorliegen eines Testreihenfolgeeffekts treffen. Um den möglichen Effekt einer variierten Testreihenfolge auf die jeweiligen Verfahren und deren Subtests zurückführen zu können, wurden sämtliche erfassten Kennwerte als abhängige Variablen in die Analysen einbezogen.

Das Versuchsdesign enthält eine unabhängige Variable mit zwei Abstufungen, dies sind die beiden Versuchsgruppen mit ihren jeweiligen Testreihenfolgen. Die verwendeten Verfahren liefern insgesamt 26 Kennwerte, diese sind die abhängigen Variablen der Auswertung. Um bei einer unabhängigen Variable und mehreren abhängigen Variablen signifikant bedeutsame Mittelwertunterschiede festzustellen, wurde eine multivariate Varianzanalyse (MANOVA) durchgeführt. Varianzanalytische Verfahren ermöglichen die Signifikanztestung von Gruppenunterschieden anhand von Mittelwertvergleichen über deren Varianzkomponenten.

3.1. Multivariate Varianzanalyse

Die MANOVA ist die methodische Erweiterung der univariaten Varianzanalyse (ANOVA) für mehr als eine abhängige Variable bei mehreren unabhängigen Variablen. Analog zur ANOVA wird das Verhältnis systematischer zu unsystematischer Varianz untersucht. Die Gesamtvarianz der beobachteten Werte wird in verschiedene Anteile zerlegt, um den systematischen Einfluss der unabhängigen Variablen auf die abhängigen Variablen vom unsystematischen Einfluss (Fehlervarianz) zu trennen. So ist es möglich, den Anteil der Varianz in den Daten zu ermitteln, der auf die experimentelle Variation der unabhängigen Variable zurückzuführen ist, und darüber hinaus das Ausmaß dieses Anteils auf statistische Signifikanz zu überprüfen.

Es wird getestet, ob über alle abhängigen Variablen die Varianz zwischen den Versuchsgruppen substanziell größer ist als die Varianz innerhalb der Gruppen. Ist die Varianz zwischen den Gruppen größer als innerhalb der Gruppen und sind diese Unterschiede im Hinblick auf ein vorher festgelegtes Signifikanzniveau bedeutsam, ist die Einteilung der Gruppen sinnvoll. Da in der vorliegenden Studie die Testreihenfolge der einzige systematische und experimentell kontrollierte Unterschied zwischen den beiden Versuchsgruppen ist, wäre dieser Faktor somit für die Gruppenunterschiede verantwortlich zu machen.

Im Gegensatz zur ANOVA werden die Abweichungsquadratsummen nicht nur einer sondern aller abhängigen Variablen und die Kreuzprodukte der gepaarten abhängigen Variablen in die Signifikanzprüfung einbezogen. Dabei wird mit zumindest drei unterschiedlichen Abweichungsquadratsummen-Kreuzprodukt Matrizen (QSKP) gearbeitet, die $QSKP_{gesamt}$, die $QSKP_{zwischen\ den\ Gruppen}$ und die

$QSKP_{innerhalb\ der\ Gruppen}$. Wobei $QSKP_{zwischen\ den\ Gruppen}$ im idealen Fall vollständig den Anteil der durch die experimentelle Variation bedingten Varianz in den Daten darstellt. Es besteht ein rechnerischer Zusammenhang zwischen den Matrizen, der in *Formel 1* dargestellt ist.

Formel 1:

Mathematischer Zusammenhang der Varianzanteile bei der Berechnung einer MANOVA

$$QSKP_{gesamt} = QSKP_{zwischen\ den\ Gruppen} + QSKP_{innerhalb\ der\ Gruppen}$$

Dividiert man die einzelnen Werte der QSKP-Matrizen durch ihre korrespondierenden Freiheitsgrade, erhält man die Varianz und Kovarianz der abhängigen Variablen. Ziel ist es nun, diese Matrizen ohne Informationsverlust auf einzelne prüfbare Werte zu reduzieren. Zu diesem Zweck bestimmt man die Determinanten dieser Matrizen, anhand derer sich unterschiedliche Prüfgrößen herleiten lassen, die approximativ *F*-verteilt sind.

SPSS stellt bei der Berechnung einer MANOVA insgesamt vier verschiedene Prüfgrößen zur Verfügung, die sich jedoch im hier vorliegenden Fall mit einer zweistufigen unabhängigen Variable in ihrer Aussagekraft und auch im resultierenden *p*-Wert nicht unterscheiden. Verwendet wird hier, wie auch in Experiment 1 und 2, die Prüfgröße *Wilks* Λ . *Wilks* Λ kann mit korrespondierenden Freiheitsgraden in einen approximativ *F*-verteilt Wert umgerechnet werden. Dieser Wert kann dann anhand eines kritischen *F*-Wertes auf Signifikanz bewertet werden. Dabei wird die Wahrscheinlichkeit der beobachteten Daten unter Annahme gleicher Populationsmittelwerte aller Gruppen gegenüber der Wahrscheinlichkeit unter Annahme, dass die Populationsmittelwerte identisch mit denen der Stichprobenmittelwerte der verschiedenen Gruppen sind, geprüft.

Für die MANOVA werden also wie oben beschrieben zunächst die Determinanten der QSKP-Matrizen gebildet und *Wilks* Λ nach *Formel 2* berechnet.

Formel 2:

Berechnung von Wilks Λ : $\det(QSKP)$ = Determinante der Quadratsummen-Kreuzprodukt-Matrizen.

$$\Lambda = \frac{|\det(QSKP)_{innerhalb\ der\ Gruppen}|}{|\det(QSKP)_{zwischen\ den\ Gruppen} + \det(QSKP)_{innerhalb\ der\ Gruppen}|}$$

Λ ist das Verhältnis der Fehlervarianz zur Gesamtvarianz. Aus *Formel 2* wird ersichtlich, dass hohe Werte für $\det(QSKP)_{\text{zwischen den Gruppen}}$ zu kleinen Werten für Λ führen. Hohe Werte für diese Determinante kommen zustande, wenn die Varianz zwischen den Gruppen groß ist. Statistische Signifikanz wird erreicht, wenn Λ klein ist.

Wilks Λ ist ebenfalls die Prüfgröße der Diskriminanzanalyse und wird auch aufgrund ihrer direkten Verbindung zur Effektstärke η^2 , denn $\eta^2 = 1 - \Lambda$, von Kubinger (2011) empfohlen. SPSS gibt als Effektstärke bei der Berechnung einer MANOVA das partielle η^2 an. Bei mehr als einer unabhängigen Variablen mit zumindest zwei Abstufungen gibt dieser Wert Aufschluss über den jeweiligen Anteil eines Haupt- oder Interaktionseffekts an der Gesamteffektstärke und erlaubt so innerexperimentelle Vergleiche. Das partielle η^2 ist vom Wert her mit dem regulären η^2 identisch, wenn, wie es in der vorliegenden Studie der Fall ist, lediglich eine zweistufige unabhängige Variable in die MANOVA einbezogen wird.

Der entscheidende Vorteil der MANOVA gegenüber den univariaten Ansätzen liegt darin, dass über die Kreuzprodukte auch die Korrelation der Variablen untereinander bei der Berechnung berücksichtigt wird. Diese Information ist unbeeinflusst durch die Transformation weiterhin in den Determinanten der QSKP-Matrizen enthalten. So erfasst die MANOVA auch komplexe Zusammenhänge und Interaktionseffekte. Die MANOVA besitzt durch die Berücksichtigung der Korrelation der abhängigen Variablen und durch die Vermeidung der Kumulierung des Risikos für den Fehler 1. Art, die bei einer alternativen Durchführung mehrfacher univariater Signifikanzprüfung an ein und derselben Stichprobe vorliegen würde, eine größere Teststärke hinsichtlich der Aufdeckung von Mittelwertunterschieden bei mehreren abhängigen Variablen. Sie ist für das experimentelle Design der vorliegenden Studie somit die Auswertungsmethode der Wahl.

Im Falle eines signifikanten Haupteffekts des Faktors Testreihenfolge war die Durchführung einer Diskriminanzanalyse vorgesehen. Ziel einer Diskriminanzanalyse ist es, jene lineare Kombination der abhängigen Variablen zu identifizieren, die maximal zwischen den Versuchsgruppen zu unterscheiden vermag. So können jene abhängigen Variablen identifiziert werden, die den größten Beitrag zur

Unterscheidbarkeit der Versuchsgruppen leisten. Diese wären dann am ehesten für die Testreihenfolgeeffekte verantwortlich. Der Vollständigkeit halber wurde trotz des in weiterer Folge zu berichtenden nicht signifikanten Haupteffekts im Anschluss an die MANOVA eine Diskriminanzanalyse durchgeführt. Deren Ergebnisse werden in dieser Arbeit nicht diskutiert, sind aber dem Anhang IV.3. zu entnehmen.

3.2. Anmerkung zu den herangezogenen Variablen und Kennwerten

Hinsichtlich der Art und Anzahl der herangezogenen Kennwerte ist für die Auswertung folgendes zu beachten. Von den verwendeten Verfahren werden zusätzliche Kennwerte ausgegeben, die vor allem bei der INSBAT theoriegeleitet zu hierarchisch übergeordneten Konstrukten verrechnet werden. Gemeint sind hier etwa die aggregierten Werte wie Gesamtintelligenz, fluide Intelligenz, Quantitatives Denken oder Gedächtnisleistung. Da diese Werte sich jedoch linear aus den einzelnen Subtestkennwerten errechnen, hat dies Auswirkungen auf die Berechnung einer MANOVA. Kommen einzelne Kennwerte, summiert in anderen Variablen, mehrfach in den Daten vor, steckt redundante Information in den Variablen. Die Berechnung der MANOVA wird dadurch ad absurdum geführt und die Ergebnisse sind nicht interpretierbar. Um diese Problematik zu vermeiden, wurde ausschließlich mit den Subtestkennwerten gerechnet, und es wurden sämtliche übergeordneten aggregierten Werte aus den Analysen ausgeschlossen. Die Auswertung wurde der Vollständigkeit halber zwar auch ausschließlich mit den aggregierten Werten durchgeführt, gelangt jedoch aufgrund der linearen Abhängigkeit der Werte nicht zu anderen Ergebnissen. Im Interesse einer detaillierten Auswertung und der Möglichkeit, einzelne Subtests und deren Kennwerte bei der Ergebnisinterpretation zu identifizieren, wurde ausschließlich mit den Subtestkennwerten gerechnet.

Eine weitere Problematik ergab sich bei der Analyse des Subtest *Verhinderung des planmäßigen Vorgehens* des BACOs. Der Kennwert *Geschwindigkeit* musste aus den multivariaten Analysen ausgeschlossen werden. Es besteht eine von den Entwicklern beabsichtigte inhaltliche und programmiertechnische Abhängigkeit der Kennwerte *Geschwindigkeit* und *Quantität Labyrinth*. Testpersonen, die das Labyrinth vollständig absolviert haben, erreichen im Kennwert *Quantität Labyrinth* einen Prozentrang von 100. Es besteht jedoch die Möglichkeit, den Subtest vorzeitig abubrechen und somit

das Ziel nicht zu erreichen. Hierbei wird der Prozentrang im Kennwert *Quantität Labyrinth* in Abhängigkeit vom Grad der Zielerreichung geringer. Der Kennwert *Geschwindigkeit* wird jedoch immer nur dann berechnet, wenn die Testperson das Labyrinth vollständig gemeistert hat, also einen Prozentrang von 100 im Kennwert *Quantität Labyrinth* erreicht hat. Bricht die Testperson vorher ab, wird der Kennwert *Geschwindigkeit* gar nicht erst berechnet und stattdessen ein fehlender Wert ausgegeben.

Bei der Berechnung des Box-M Tests zur Überprüfung der Homogenität der Varianz-Kovarianz-Matrizen kommt es dadurch zu einer Fehlermeldung und einem Abbruch der Berechnung. Bei dem Versuch, die Varianzen der beobachteten Werte miteinander zu korrelieren, kann aufgrund der fehlenden Werte im Kennwert *Geschwindigkeit* nur dort eine Korrelation berechnet werden, wo auch ein Wert verfügbar ist. Ein Wert für die *Geschwindigkeit* ist aber nur dann verfügbar, wenn die Testperson das Labyrinth erfolgreich beendet hat und somit im Kennwert *Quantität Labyrinth* den Prozentrang 100 erreicht hat. Da dieser dann jedoch für alle Personen identisch ist - die Abbrecher werden aufgrund fehlender Werte im Kennwert *Geschwindigkeit* nicht berücksichtigt - wird dieser zu einer Konstanten in der Berechnung. Der Kennwert *Geschwindigkeit* variiert zwar zwischen den in diesem Subtest erfolgreichen Testpersonen, kann aber nicht mit der Konstanten (Prozentrang 100 im Kennwert *Quantität Labyrinth*) korreliert werden. Die Berechnung einer Korrelation zwischen einer Variablen und einer Konstanten ist nicht möglich, die Berechnung scheitert also und mündet in einer Fehlermeldung.

Aufgrund dieser Umstände ist eine Berechnung des Box-M Tests unter Einbezug aller 26 Variablen nicht möglich, und die Interpretation der Ergebnisse der MANOVA unter gemeinsamer Berücksichtigung der Variablen *Quantität Labyrinth* und *Geschwindigkeit* ist nicht zulässig. Zur Lösung dieses Problems wurde aufgrund der unvollständigen Daten im Kennwert *Geschwindigkeit* diese Variable aus den multivariaten Analysen ausgeschlossen und separat mittels des parameterfreien *Welch*-Tests untersucht.

Leider war es auch nach Rücksprache mit den Programmierern nicht möglich, für die in diesem Subtest nicht erfolgreichen Testpersonen ebenfalls einen Geschwindigkeitskennwert zu errechnen. Im Gegensatz zur in Experiment 1 und 2 verwendeten Vorgängerversion BaCO-D ist diese Funktion in der hier verwendeten

neueren Version BACO aufgrund konzeptioneller Weiterentwicklung programmiertechnisch nicht umgesetzt, da der Geschwindigkeitskennwert bei vorzeitigem Abbruch des Subtests inhaltlich nicht sinnvoll zu interpretieren ist.

4. Ergebnisse

Die Testergebnisse und alle relevanten verfahrensspezifischen und demographischen Variablen wurden mithilfe der Exportfunktion des WTS in ein Excel-File (.csv) übertragen und weiterverwendet. Die Rohwerte wurden zu einheitlich charakterisierten Normstichproben ins Verhältnis gesetzt und die so entstehenden Prozenträge für die Rückmeldungen verwendet. Für die Analysen der vorliegenden Arbeit wurden diese Prozenträge analog zu Experiment 1 und 2 z-transformiert. Die so entstandenen flächentransformierten z-Werte vereinheitlichen die Datengrundlage und erreichen dabei zusätzlich eine bessere approximative Normalverteilung.

4.1. Deskriptive Statistik

In *Tabelle 4* werden die Mittelwerte und Standardabweichungen der mit den verwendeten Verfahren erhobenen und in die Analysen einbezogenen Kennwerte für beide Versuchsgruppen dargestellt. Die Werte ergeben sich aus der beschriebenen Flächentransformation der Rohwerte. Die Syntax, nach der diese Transformation durchgeführt wurde, ist Anhang IV.2. zu entnehmen.

Tabelle 4:
Mittelwerte und Standardabweichungen aller Subtestkennwerte in beiden
Versuchsgruppen – z-transformiert

Abhängige Variablen		Gruppe 1 LT → OPT		Gruppe 2 OPT → LT	
		<i>Mittelwert</i>	<i>SD</i>	<i>Mittelwert</i>	<i>SD</i>
Allgemeine Intelligenz	AMT	,808	,903	,747	,788
Arithmetische Kompetenz	INSBAT	-,167	1,005	,149	,814
Arithmetische Flexibilität	INSBAT	,089	,630	,134	,797
Verbal deduktives Denken	INSBAT	,642	,974	,587	,881
Verbales Kurzzeitgedächtnis	INSBAT	,702	1,060	,628	1,088
Visuelles Kurzzeitgedächtnis	INSBAT	,761	,988	,621	,676
Numerisch induktives Denken	INSBAT	,551	1,113	,385	1,173
Wortbedeutungen	INSBAT	-,008	,758	-,149	1,079
Exaktheit	AHA	,320	1,095	,532	1,007
Entschlussfreudigkeit	AHA	-,341	,930	-,228	,847
Reflexivität	AHA	,346	,875	,240	,881
Leistungsniveau	AHA	,864	,736	,853	,514
Anspruchsniveau	AHA	,334	,591	,289	,661
Frustrationstoleranz	AHA	-,035	,597	,059	,741
Zieldiskrepanz	AHA	-,185	,870	-,056	,736
Quantität Labyrinth	BACO	3,140	1,651	3,246	1,566
Geschwindigkeit	BACO	,927	,739	,922	,791
Verharren	BACO	-,249	,431	-,238	,458
Herumirren	BACO	-,071	1,061	-,123	,836
Quantität Hauptaufgabe	BACO	,218	,685	,260	,708
Quantität Nebenaufgabe	BACO	,578	,684	,552	,645
Ausgewogenheit	BACO	-,244	,561	-,241	,550
Effizienz Nebenaufgabe	BACO	,480	1,050	,549	,890
Durchhaltevermögen	BACO	,288	,875	,411	,881
Vermittelnd	ILICA	,104	,752	,012	,944
Ablenkbar	ILICA	-,261	,722	-,219	,820
Zielorientiert	ILICA	,548	,953	,667	,802

Die vom WTS geschätzte Durchführungsdauer der Testbatterie ohne Berücksichtigung der schätzungsweise 20 bis 30 Minuten für den ILICA lag bei etwa 200 Minuten. Aufgrund der fehlenden Daten zur Bearbeitungszeit des ILICAs, wird dieser bei der folgenden Darstellung der erfassten Gesamttestdauer nicht berücksichtigt. Tatsächlich betrug die Bearbeitungszeit ohne Berücksichtigung des ILICAs im Mittel etwa 190 Minuten. *Tabelle 5* enthält eine ausführliche Übersicht über Mittelwerte und Standardabweichungen der Testdauer, aufgeschlüsselt nach den einzelnen Verfahren und für beide Versuchsgruppen.

Tabelle 5:

Darstellung der Mittelwerte und Standardabweichungen der Testdauer insgesamt und aufgeschlüsselt nach Versuchsgruppen und Verfahren

Test	Gesamt		Gruppe 1 LT → OPT		Gruppe 2 OPT → LT	
	<i>Mittelwert</i>	<i>SD</i>	<i>Mittelwert</i>	<i>SD</i>	<i>Mittelwert</i>	<i>SD</i>
AMT	39,06	15,242	42,43	14,753	35,69	15,179
INSBAT	114,57	23,018	117,94	26,952	111,20	18,039
AHA	19,01	3,151	18,23	3,422	19,80	2,677
BACO	16,66	2,933	17,03	3,276	16,29	2,539
Gesamt	189,30	37,069	195,63	40,460	182,97	32,699

4.2. Voraussetzungsprüfungen der MANOVA

Die Berechnung einer MANOVA setzt einige Annahmen voraus, die im Vorfeld der Analyse überprüft werden müssen. Um die Mittelwerte der Versuchsgruppen vergleichend zu analysieren, ist vorauszusetzen, dass sich gegebenenfalls auch nur diese voneinander unterscheiden und nicht etwa die Varianzen und Kovarianzen anderweitig voneinander abweichen. Die Voraussetzungsprüfungen wurden mit adäquaten Verfahren durchgeführt, die Vorgehensweise und die Ergebnisse sind im Folgenden dargestellt und erläutert.

Die MANOVA stellt einen Spezialfall des Allgemeinen Linearen Modells dar, und ist nur für die Analyse linearer Zusammenhänge geeignet, da mit der zugrundeliegenden Korrelation nach Pearson nur lineare Zusammenhänge erfasst und korrekt dargestellt werden. Neben der Linearität des Zusammenhanges ist für die Berechnung einer MANOVA die multivariate Normalverteilung für jede Zelle innerhalb aller Versuchsgruppen vorauszusetzen. Die Homogenität der Varianz-Kovarianz-Matrizen muss ebenfalls geprüft werden, die Überprüfung der Gleichheit der Fehlervarianzen ist als optional zu bezeichnen und nur für folgende univariate Analysen tatsächlich notwendig.

4.2.1. Linearität

Die Linearität oder der lineare Zusammenhang zwischen den abhängigen Variablen ist eine grundlegende Voraussetzung für die Berechnung einer MANOVA, da die zugrundeliegenden Formeln nur für lineare und nicht für beispielsweise kurvilineare Zusammenhänge vorgesehen sind. Diese Linearitätsannahme wird mithilfe von Streudiagrammen der paarweise aufgetragenen Variablen überprüft, wobei jede Kombination einzeln grafisch überprüft werden muss. Ist das Streudiagramm ovalförmig um eine imaginäre Gerade angeordnet, ist dies ein Hinweis auf einen linearen Zusammenhang. Da diese Analysen sehr umfangreich sind, wird auf eine vollständige Darstellung verzichtet, und es wird nur exemplarisch der Zusammenhang zwischen den Kennwerten *Allgemeine Intelligenz* des AMTs und *Verbal-Deduktives Denken* aus der INSBAT in *Abbildung 2* dargestellt. Ein Screening aller Zusammenhänge wurde vorgenommen. Es konnten keine Auffälligkeiten, die eine Voraussetzungsverletzung vermuten lassen, festgestellt werden, da ein beobachtbarer Trend eines linearen Zusammenhangs bei so gut wie allen Variablenkombinationen vorhanden war. Somit wird die Linearität der Zusammenhänge als gegeben angenommen.

Der lineare Zusammenhang ist anhand von *Abbildung 2* exemplarisch gut zu erkennen, die Kennwerte zeigen eine oval-förmige Streuung um eine imaginäre Ursprungsgerade.

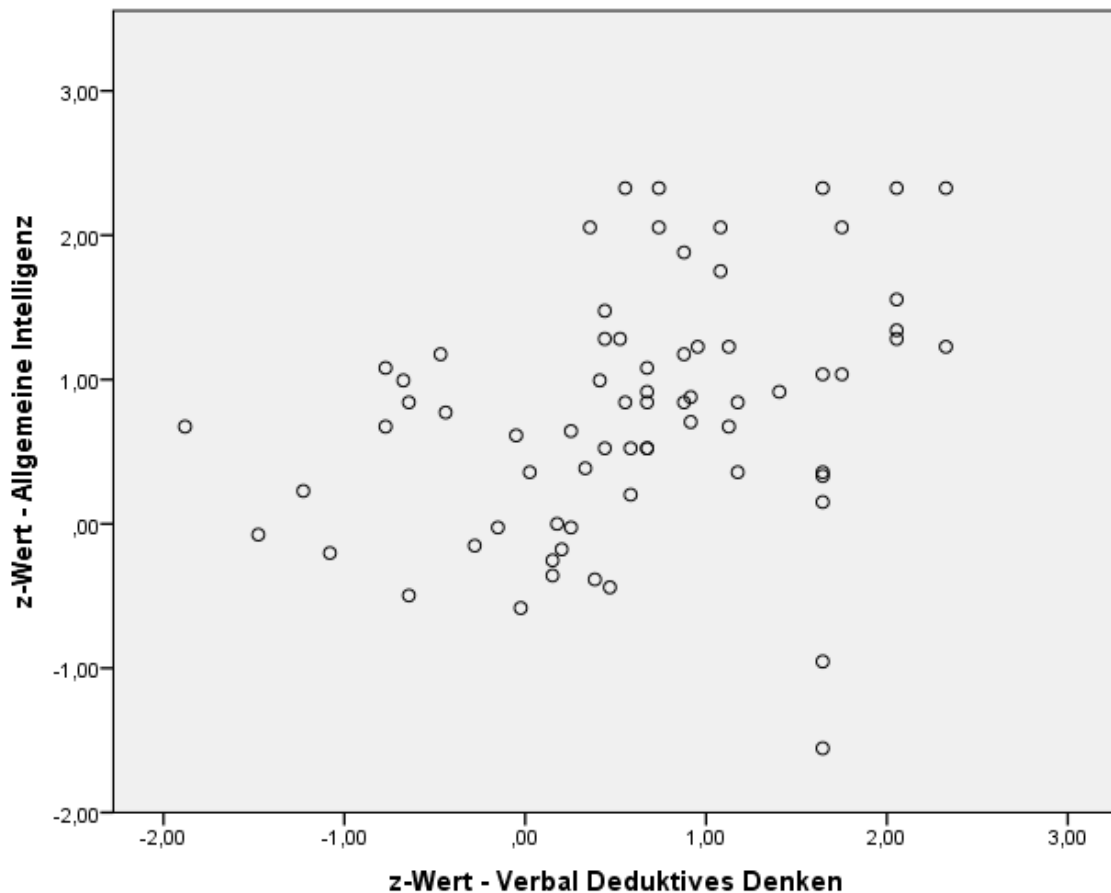


Abbildung 2: Diagramm des Zusammenhangs der z-Werte der Variablen Allgemeine Intelligenz (AMT) und Verbal Deduktives Denken (INSBAT)

4.2.2. Multivariate Normalverteilung

Multivariate Normalverteilung bedeutet, dass alle abhängigen Variablen und sämtliche linearen Kombinationen derselben normalverteilt sind. Ist diese Voraussetzung erfüllt, sind auch die Residuen normalverteilt und die Berechnung einer MANOVA ist zulässig (Tabachnick, 2007).

Obwohl die Prüfung der Normalverteilung eine Voraussetzung zur Durchführung einer MANOVA darstellt (Tabachnick, 2007), finden sich auch Studien, die eine Verletzung dieser Annahme ab gewissen Stichprobenumfängen als in ihren Auswirkungen vernachlässigbar ansehen (Rasch & Guiard, 2004). Die Überprüfung der Normalverteilung wurde hier aufgrund dieser uneindeutigen Situation trotzdem durchgeführt.

Die Überprüfung der multivariaten Normalverteilung ist in Ermangelung geeigneter Verfahren bisher auf direktem Wege nicht zufriedenstellend möglich, da eine sehr große Anzahl linearer Kombinationen auf ihre Verteilungen überprüft werden müsste. Die verfügbaren Verfahren kommen dabei zwangsläufig aufgrund der steigenden Teststärke bei wiederholten Testungen zu fälschlicherweise signifikanten Ergebnissen. Die verfügbaren Methoden basieren größtenteils auf multivariaten Weiterentwicklungen des Shapiro Wilk Tests für univariate Normalverteilung und liefern keine global aussagekräftigen Werte. In Hinblick auf die Robustheit der MANOVA bei geringfügiger Verletzung der multivariaten Normalverteilungsvoraussetzung (Rasch & Guiard, 2004) wurde stattdessen, wie in der Praxis üblich, lediglich auf univariate Normalverteilung der Kennwerte getestet. Diese stellt ihrerseits eine Voraussetzung für die multivariate Normalverteilung dar. Generell ist zu sagen, dass auch univariat normalverteilte Daten zu einer besseren Interpretierbarkeit der Ergebnisse einer MANOVA beitragen.

Die univariate Normalverteilung kann entweder grafisch anhand der Schiefe und der Kurtosis der Verteilung oder statistisch mithilfe geeigneter Verfahren überprüft werden. Die vorliegenden Daten wurden mit zwei unterschiedlichen Verfahren statistisch auf ihre Normalverteilung überprüft. Die herangezogenen Prüfgrößen nach Kolmogorov-Smirnov und Shapiro-Wilk unterscheiden sich hinsichtlich ihrer Sensibilität für unterschiedliche Arten von Normalverteilungsverletzungen. Beide Verfahren sind am besten komplementär anzuwenden und die Ergebnisse gemeinsam in die Entscheidung einzubeziehen. Die Ergebnisse der Testung auf univariate Normalverteilung aller 26 in die Berechnungen einbezogenen Variablen sind in *Tabelle 6* dargestellt. Die Analysen erfolgten für alle Kennwerte mit beiden Prüfverfahren nach Gruppen getrennt, dies entspricht $26 \times 2 \times 2 = 104$ Zellen. Farblich markiert sind jene Werte, die signifikant von der Normalverteilungsannahme abweichen.

Tabelle 6:

Ergebnisse der Testung auf univariate Normalverteilung nach Kolmogorov-Smirnov und Shapiro-Wilk

Tests auf Normalverteilung							
Abhängige Variable	Gruppe	Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistik	df	Signifikanz	Statistik	df	Signifikanz
Allgemeine Intelligenz	1	,100	35	,200	,964	35	,298
	2	,093	35	,200	,968	35	,390
Arithmetische Kompetenz	1	,138	35	,088	,970	35	,440
	2	,100	35	,200	,975	35	,579
Arithmetische Flexibilität	1	,182	35	,005	,917	35	,011
	2	,232	35	,000	,831	35	,000
Verbal deduktives Denken	1	,134	35	,114	,960	35	,225
	2	,109	35	,200	,975	35	,578
Verbales Kurzzeitgedächtnis	1	,132	35	,129	,974	35	,558
	2	,103	35	,200	,968	35	,396
Visuelles Kurzzeitgedächtnis	1	,195	35	,002	,866	35	,001
	2	,207	35	,001	,877	35	,001
Numerisch induktives Denken	1	,128	35	,154	,936	35	,041
	2	,126	35	,178	,946	35	,085
Wortbedeutungen	1	,168	35	,014	,951	35	,120
	2	,127	35	,165	,941	35	,058
Exaktheit	1	,085	35	,200	,973	35	,516
	2	,090	35	,200	,971	35	,467
Entschlussfreudigkeit	1	,130	35	,141	,944	35	,072
	2	,117	35	,200	,944	35	,072
Reflexivität	1	,094	35	,200	,973	35	,541
	2	,164	35	,018	,954	35	,152
Leistungsniveau	1	,085	35	,200	,975	35	,590
	2	,173	35	,010	,945	35	,080

Tests auf Normalverteilung

Abhängige Variable	Gruppe	Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistik	df	Signifikanz	Statistik	df	Signifikanz
Anspruchsniveau	1	,099	35	,200	,972	35	,501
	2	,106	35	,200	,964	35	,304
Frustrationstoleranz	1	,157	35	,029	,934	35	,036
	2	,083	35	,200	,978	35	,691
Zieldiskrepanz	1	,201	35	,001	,896	35	,003
	2	,130	35	,141	,959	35	,212
Quantität Labyrinth	1	,524	35	,000	,376	35	,000
	2	,533	35	,000	,318	35	,000
Verharren	1	,500	35	,000	,479	35	,000
	2	,499	35	,000	,481	35	,000
Herumirren	1	,170	35	,012	,894	35	,003
	2	,085	35	,200	,981	35	,804
Quantität Hauptaufgabe	1	,110	35	,200	,969	35	,416
	2	,089	35	,200	,991	35	,989
Quantität Nebenaufgabe	1	,140	35	,081	,951	35	,124
	2	,177	35	,007	,910	35	,008
Ausgewogenheit	1	,130	35	,143	,965	35	,317
	2	,111	35	,200	,981	35	,801
Effizienz Nebenaufgabe	1	,141	35	,078	,925	35	,019
	2	,179	35	,006	,884	35	,002
Durchhaltevermögen	1	,110	35	,200	,945	35	,080
	2	,070	35	,200	,985	35	,915
Vermittelnd	1	,160	35	,023	,951	35	,119
	2	,109	35	,200	,975	35	,586
Ablenkbar	1	,147	35	,054	,957	35	,184
	2	,137	35	,096	,937	35	,046
Zielorientiert	1	,150	35	,044	,936	35	,043
	2	,123	35	,200	,961	35	,252

Es zeigt sich, dass nach Shapiro Wilk die Verteilungen lediglich in 17 von 52 Zellen signifikant nicht normalverteilt sind. Nach Kolmogorov-Smirnov erfüllen die Verteilungen in 18 von 52 Zellen die Normalverteilungsvoraussetzung nicht.

Bei gleicher Zellbesetzung in gruppierten Daten ist der *F*-Test für univariate ANOVAs nach Mardia (1971) bei zumindest 20 Fehlerfreiheitsgraden bereits als robust gegenüber Verletzungen der Normalverteilungsvoraussetzungen anzunehmen. Seo, Kanda, und Fujikoshi (1995) belegten in Simulationsstudien bereits ab $N = 40$ die Robustheit der MANOVA gegenüber Verletzungen der Normalverteilungsannahme. Durch die in Abschnitt III.4. dargestellte Flächentransformation der Kennwerte konnte eine approximativ bessere Normalverteilung der Daten erreicht werden. Die Normalverteilungsprüfung z-transformierter Werte ist eine Prüfung der Mediane und nicht der Mittelwerte der Gruppen und führt zu näherungsweise besseren Ergebnissen (Tabachnick, 2007). Da die multivariate Varianzanalyse wie vorausgehend dargestellt auch bei Verletzung der Normalverteilungsvoraussetzung ein robustes Analyseverfahren bleibt, sind die anhand der Shapiro Wilk und Kolmogorov-Smirnov Tests festgestellten Voraussetzungsverletzungen als nicht gravierend anzusehen.

4.2.3. Gleichheit der Fehlervarianzen

Obwohl die Gleichheit der Fehlervarianzen von SPSS standardmäßig bei der Durchführung einer MANOVA überprüft wird, stellt sie keine notwendige Voraussetzung für die legitime Berechnung einer MANOVA dar. Die Gleichheit der Fehlervarianzen ist Voraussetzung für univariate Analysen und ist somit für die in der Regel der MANOVA folgenden ANOVAs relevant. Sie wird mit dem Levene-Test überprüft und getestet zusammengekommen mit dem Box-M Test die Voraussetzung der Homogenität der Varianz-Kovarianz Matrizen. Es ist darauf hinzuweisen, dass mit dem Levene-Test lediglich die Homogenität der Fehlervarianzen und nicht die wesentlich relevantere Homogenität der Kovarianzen überprüft wird. Der Levene-Test wurde an dieser Stelle durchgeführt, weil er im Vergleich zum Box-M Test robuster gegen Verletzungen der (multivariaten) Normalverteilung (Tabachnick, 2007) und notwendige Voraussetzung für die folgenden univariaten Auswertungen ist.

Die Gleichheit der Fehlervarianzen wurde mithilfe des Levene-Tests überprüft, die Ergebnisse sind *Tabelle 7* zu entnehmen.

Tabelle 7:

Ergebnisse des Levene-Tests für sämtliche Variablen

Levene-Test auf Gleichheit der Fehlervarianzen					
Abhängige Variable	Test	<i>F</i>	df1	df2	Signifikanz
Allgemeine Intelligenz	AMT	,171	1	68	,680
Arithmetische Kompetenz	INSBAT	,333	1	68	,566
Arithmetische Flexibilität	INSBAT	2,591	1	68	,112
Verbal deduktives Denken	INSBAT	1,650	1	68	,203
Verbales Kurzzeitgedächtnis	INSBAT	,240	1	68	,626
Visuelles Kurzzeitgedächtnis	INSBAT	1,516	1	68	,223
Numerisch induktives Denken	INSBAT	,189	1	68	,665
Wortbedeutungen	INSBAT	3,088	1	68	,083
Exaktheit	AHA	,225	1	68	,637
Entschlussfreudigkeit	AHA	1,062	1	68	,307
Reflexivität	AHA	,009	1	68	,924
Leistungsniveau	AHA	1,934	1	68	,169
Anspruchsniveau	AHA	,277	1	68	,601
Frustrationstoleranz	AHA	3,983	1	68	,050
Zieldiskrepanz	AHA	1,079	1	68	,303
Quantität Labyrinth	BACO	,307	1	68	,581
Verharren	BACO	,065	1	68	,800
Herumirren	BACO	,240	1	68	,625
Quantität Hauptaufgabe	BACO	,001	1	68	,979
Quantität Nebenaufgabe	BACO	,775	1	68	,382
Ausgewogenheit	BACO	,023	1	68	,881
Effizienz Nebenaufgabe	BACO	1,550	1	68	,217
Durchhaltevermögen	BACO	,001	1	68	,974
Vermittelnd	ILICA	,664	1	68	,418
Ablenkbar	ILICA	,063	1	68	,803
Zielorientiert	ILICA	2,271	1	68	,136

Hinsichtlich der Voraussetzung der Gleichheit der Fehlervarianzen zeigen sich alle herangezogenen Kennwerte als nicht signifikant und damit unauffällig.

Obwohl aufgrund des in weiterer Folge behandelten nicht signifikanten Box-M Tests von der Gleichheit der Fehlervarianzen ausgegangen werden kann und der nicht signifikante Haupteffekt der Testreihenfolge eine univariate Analyse der Testkennwerte überflüssig macht, wurden hier der Vollständigkeit halber die Ergebnisse des Levene-Test auf Gleichheit der Fehlervarianzen angegeben. Diese bestätigen, dass eine Interpretation der MANOVA Ergebnisse zulässig ist und auch die univariaten Analysen bei einem signifikanten Haupteffekt sinnvoll interpretierbar wären.

4.2.4. Homogenität der Varianz-Kovarianz-Matrizen

Die Homogenität der Varianz-Kovarianz-Matrizen ist eine zentrale Voraussetzung für die Berechnung einer MANOVA. Zur Überprüfung dieser Annahme wird der Box-M Test verwendet, welcher die multivariate Ausführung der für univariate Analysen notwendigen Prüfung der Varianzhomogenität darstellt. Er testet die Annahme, dass die beobachteten Varianzen und Kovarianzen in den QSKP-Matrizen über alle Versuchsgruppen homogen sind.

Der Box-M Test ist ein sehr sensibler Test, da er das Risiko 1. Art bei Abweichungen von der multivariaten Normalverteilung nicht annähernd einhält und aufgrund der hohen Teststärke schnell signifikant wird. Laut Tabachnick (2007) hat ein signifikanter Box-M Test bei gleicher Gruppengröße und Zellbesetzung allerdings keinen großen Einfluss auf die Robustheit der MANOVA.

Das Ergebnis des Box-M Test auf Gleichheit der Kovarianzmatrizen ist der *Tabelle 8* zu entnehmen.

Tabelle 8:
Ergebnisse des Box-M-Tests
Box-Test auf Gleichheit der
Kovarianzenmatrizen

Box-M-Test	634,719
F	1,059
df1	351
df2	14025,575
Signifikanz	,216

Der Box-M Test ist nicht signifikant ($p = 0,216$), und die Homogenität der Varianz-Kovarianz-Matrizen kann somit unter Einbezug der Ergebnisse des Levene-Tests als gegeben angenommen werden. Die Anwendung einer multivariaten Varianzanalyse ist somit berechtigt, die resultierenden *F*-Werte verlässlich und die Ergebnisse können berechtigterweise interpretiert werden.

4.3. Ergebnisse der inferenzstatistischen Analysen

Im Anschluss an die Voraussetzungsprüfungen wurde eine MANOVA zur Signifikanzprüfung möglicher Mittelwertunterschieden zwischen den Versuchsgruppen durchgeführt. Die Fragestellung, operationalisiert durch die Hypothesen 1 und 2, konnte so überprüft und beantwortet werden.

Die Hypothesen 1 und 2 sind anhand der durchgeführten Analysen eindeutig zu beantworten. Die Ergebnisse sind in *Tabelle 9* dargestellt und geben keinen Anlass zur Annahme der Alternativhypothesen H1.1 und H2.1, sondern sprechen eindeutig für die Beibehaltung der H1.0 und H2.0. Der Faktor *Gruppe*, welcher die beiden Versuchsgruppen hinsichtlich der bearbeiteten Testreihenfolge aufteilt, hat keinen Einfluss auf die Testleistung der Teilnehmer. Es konnte somit kein signifikanter

Haupteffekt der Testreihenfolge gefunden werden ($p = 0,998$; $F = 0,323$; $\eta^2 = 0,163$). Die Ergebnisse der multivariaten Varianzanalyse sind der *Tabelle 9* zu entnehmen.

Tabelle 9:
Ergebnisse der MANOVA

Multivariate Tests						
Teststatistik	Wert	<i>F</i>	df Hypothese	df Fehler	Signifikanz	η^2
Pillai-Spur	,163	,323	26	43	,998	,163
<i>Wilks Λ</i>	,837	,323	26	43	,998	,163
Hotelling-Spur	,195	,323	26	43	,998	,163
Größte Wurzel nach Roy	,195	,323	26	43	,998	,163

Die folgenden univariaten Analysen untermauern das Ergebnis eines nicht signifikanten Einflusses der Testreihenfolgenvariation auf Ebene der einzelnen Kennwerte. Wie der *Tabelle 10* zu entnehmen ist, kann auch hier in keinem Kennwert ein signifikanter Gruppenunterschied festgestellt werden. Somit ist auch auf univariater Analyseebene, unter Berücksichtigung der Bonferroni-Korrektur zur Adjustierung der Kumulierung des Fehlers 1. Art bei wiederholtem Testen an ein und derselben Stichprobe, kein Effekt der Testreihenfolge zu beobachten ($0,981 > p > 0,151$).

Tabelle 10:
Ergebnisse der ANOVAs

Univariate Tests							
Abhängige Variable	Test	SS	df	MS	F	Sig.	η^2
Allgemeine Intelligenz	AMT	,066	1	,066	,092	,763	,001
Arithmetische Kompetenz	INSBAT	,054	1	,054	,062	,803	,001
Arithmetische Flexibilität	INSBAT	1,766	1	1,766	2,112	,151	,030
Verbal deduktives Denken	INSBAT	,035	1	,035	,067	,796	,001
Verbales Kurzzeitgedächtnis	INSBAT	,094	1	,094	,082	,776	,001
Visuelles Kurzzeitgedächtnis	INSBAT	,341	1	,341	,477	,492	,007
Numerisch induktives Denken	INSBAT	,482	1	,482	,368	,546	,005
Wortbedeutungen	INSBAT	,348	1	,348	,401	,529	,006
Exaktheit	AHA	,788	1	,788	,712	,402	,010
Entschlussfreudigkeit	AHA	,224	1	,224	,283	,596	,004
Reflexivität	AHA	,195	1	,195	,253	,616	,004
Leistungsniveau	AHA	,002	1	,002	,006	,940	,000
Anspruchsniveau	AHA	,036	1	,036	,091	,763	,001
Frustrationstoleranz	AHA	,154	1	,154	,339	,562	,005
Zieldiskrepanz	AHA	,289	1	,289	,446	,507	,007
Quantität Labyrinth	BACO	,217	1	,217	,084	,773	,001
Verharren	BACO	,002	1	,002	,010	,919	,000
Herumirren	BACO	,048	1	,048	,053	,819	,001
Quantität Hauptaufgabe	BACO	,031	1	,031	,064	,802	,001
Quantität Nebenaufgabe	BACO	,011	1	,011	,026	,873	,000
Ausgewogenheit	BACO	,000	1	,000	,001	,981	,000
Effizienz Nebenaufgabe	BACO	,083	1	,083	,087	,769	,001
Durchhaltevermögen	BACO	,263	1	,263	,341	,561	,005
Vermittelnd	ILICA	,148	1	,148	,204	,653	,003
Ablenkbar	ILICA	,031	1	,031	,051	,821	,001
Zielorientiert	ILICA	,247	1	,247	,319	,574	,005

Der aufgrund der in Abschnitt III.3.2. beschriebenen Komplikationen aus den Analysen ausgeschlossene Kennwert *Geschwindigkeit* des BACOs wurde separat einer univariaten Analyse unterzogen und mittels *Welch*-Test auf Gruppenunterschiede überprüft. Das Ergebnis dieser Testung ist *Tabelle 11* zu entnehmen. Der *Welch*-Test ergab für den Kennwert *Geschwindigkeit* ein nicht signifikantes Ergebnis und lässt auch in diesem Wert keinen Effekt der Testreihenfolge erkennen ($p = 0,972$; $F = 0,001$).

Tabelle 11:
Ergebnis des *Welch*-Tests

Teststatistik	Wert	df1	df2	Signifikanz
<i>Welch</i> -Test	,001	1	59,720	,972

5. Interpretation

Die Ergebnisse der inferenzstatistischen Analysen ergeben ein einheitliches Bild. Es konnte weder ein signifikanter Haupteffekt der unabhängigen Variable *Testreihenfolge* noch für irgendeinen der erhobenen Kennwert ein signifikanter Gruppenunterschied festgestellt werden. Aufgrund der durchweg nicht signifikanten Ergebnisse lässt sich sagen, dass in der vorliegenden Studie die experimentelle Variation der Testreihenfolge keinen Effekt auf die Testleistung der Teilnehmer hatte.

Zur Rekapitulation sind die Hypothesen aus Abschnitt III.1.1 hier erneut aufgeführt.

Hypothese 1 Reihenfolgeeffekte LT → OPT:

H1.0: Die Bearbeitung kognitiver Leistungstests beeinflusst die Bearbeitung darauf folgender Objektiver Persönlichkeitstests nicht.

H1.1: Die Bearbeitung kognitiver Leistungstests beeinflusst die Bearbeitung darauf folgender Objektiver Persönlichkeitstests.

Hypothese 2 Reihenfolgeeffekte OPT → LT:

H2.0: Die Bearbeitung Objektiver Persönlichkeitstests beeinflusst die Bearbeitung darauf folgender kognitiver Leistungstests nicht.

H2.1: Die Bearbeitung Objektiver Persönlichkeitstests beeinflusst die Bearbeitung darauf folgender kognitiver Leistungstests.

Hypothese 3 Testreihenfolgeeffekte bei Vorgabe des AMTs:

H3.0: Die Vorgabe des AMTs hat keinen Einfluss auf das Auftreten von Testreihenfolgeeffekten.

H3.1: Die Vorgabe des AMTs hat einen Einfluss auf das Auftreten von Testreihenfolgeeffekten.

Es konnten keine Belege für die vermutete gegenseitige Beeinflussung der Testblöcke gefunden werden. Die Alternativhypothesen H1.1 und H2.1 konnten anhand der hier erhobenen Daten einer statistischen Überprüfung nicht standhalten und mussten zugunsten der Nullhypothesen H1.0 und H2.0 abgelehnt werden. Die Bearbeitung der kognitiven Leistungstests beeinflusste die darauf folgende

Bearbeitung der Objektiven Persönlichkeitstests nicht. Auch bei umgekehrter Reihenfolge, also bei Bearbeitung der Objektiven Persönlichkeitstests mit anschließender Bearbeitung der kognitiven Leistungstests, konnten keine Effekte festgestellt werden. Ein ergebnisrelevanter Effekt von Testreihenfolgenvariationen in Auswahlprozessen ist aufgrund dieser Ergebnisse zunächst nicht anzunehmen.

Die Auswertungen kamen zu ausschließlich nicht signifikanten Ergebnissen. Auch lassen sich bei genauer Betrachtung der Mittelwerte der abhängigen Variablen in *Tabelle 4* keine Tendenzen in Richtung eines möglichen Reihenfolgeeffektes feststellen. Die Gruppenmittelwerte unterscheiden sich lediglich marginal.

In der vorliegenden Studie konnte der in Experiment 1 gefundene Haupteffekt einer Testreihenfolgenvariation nicht repliziert werden. Auch die dort berichteten Auffälligkeiten in den Kennwerten der AHA blieben aus. Die in Experiment 2 zwischen den Gruppen bedeutsam abweichenden Kennwerte zur Gedächtnisleistung erwiesen sich in Experiment 3 ebenfalls als unauffällig.

Da die angestrebte Vergleichbarkeit der Stichproben aus Experiment 2 und 3 wie in Abschnitt III.2.4. dargestellt nicht gelungen ist, sind die hier dargestellten Ergebnisse für die herangezogene Stichprobe spezifisch gültig und eignen sich nur bedingt zur Verallgemeinerung. Es wird jedoch deutlich, dass bei gleichem Versuchsdesign, unter sehr ähnlichen Versuchsbedingungen und bei Verwendung einer beinahe identischen Testbatterie andere Ergebnisse gefunden wurden, als in Experiment 1 und 2. Hypothese 3 ist zwar in Bezug auf die vorliegende Studie eindeutig zugunsten der Nullhypothese H3.0 zu beantworten, aber ein Vergleich mit Experiment 2 scheitert und kann somit die damit verknüpfte zweite Forschungsfrage aus Abschnitt III.1. nicht eindeutig aufklären. In der vorliegenden Arbeit hat die Vorgabe des AMTs nicht zu Testreihenfolgeneffekten geführt, womit dieser Matrizentest nicht für Testreihenfolgeneffekte verantwortlich gemacht werden kann, allerdings gelingt es nicht die Ergebnisse aus Experiment 1 zu erklären.

Somit bleibt eine Auseinandersetzung mit Fragestellung drei aus Abschnitt III.1. ausstehend, die neben der Verfahrensauswahl vor allem die Stichprobenspezifität von Testreihenfolgeneffekten thematisiert. Es wurden hierzu keine statistisch prüfbaren Hypothesen formuliert, durch die folgende Diskussion und der dort vorgenommenen weiterführenden inhaltlichen Interpretation der hier gefundenen Ergebnisse sowie unter Berücksichtigung der im Theorieteil ausgearbeiteten Erkenntnisse zu

Reihenfolgeeffekten kann trotzdem zur weiteren Aufklärung dieser Fragestellungen beigetragen werden.

6. Diskussion

Betrachtet man die vorliegende Arbeit für sich allein genommen, können die hier gefundenen empirischen Ergebnisse in Hinblick auf die formulierten Hypothesen eindeutig interpretiert werden. Es wird jedoch schnell deutlich, dass unter Berücksichtigung der Vorgängerstudien aufgrund der eingeschränkten Vergleichbarkeit der Stichproben und der Variation der Testauswahl die aufgeworfenen Forschungsfragen nicht eindeutig geklärt werden können. Der vorab erhobene Anspruch einer Vergleichbarkeit der Stichproben konnte nicht gewährleistet werden und schränkt so die Generalisierbarkeit der hier gefundenen Ergebnisse ein.

Dieser Umstand macht es notwendig, diese Arbeit auf drei Stufen zu diskutieren. Auf unterster Stufe steht die alleinige Betrachtung der bereits dargestellten nicht signifikanten Ergebnisse der Hypothesentestung in dieser Studie. Für sich genommen, sprechen diese unter Berücksichtigung der eingeschränkten Generalisierbarkeit aufgrund der hier herangezogenen Stichprobe gegen das Auftreten von Testreihenfolgeeffekten in Personalauswahlsituationen. Es stellt sich allerdings die Frage, warum es hier nicht zu Testreihenfolgeeffekten gekommen ist, wo es doch bereits in vergleichbaren Studien gegenteilige Ergebnisse gegeben hat.

Bezieht man auf einer zweiten Stufe die Erkenntnisse aus den Experimenten 1 und 2 mit ein, zeigt sich ein wenig einheitliches Bild. In Experiment 1 lag ein Reihenfolgeeffekt als Haupteffekt vor, der sich in Kennwerten der OPTs zeigte. In Experiment 2 konnte kein Haupteffekt der Testreihenfolge gefunden werden. In Experiment 3 konnte hingegen weder ein Haupteffekt der Testreihenfolge gefunden werden, noch zeigten sich Auffälligkeiten in irgendeinem der herangezogenen Testkennwerte. Die Ergebnisse der Studien widersprechen sich auf den ersten Blick, so dass offenkundig weiterführende Überlegungen angestellt werden müssen.

Eine detailliertere Betrachtung aller drei Studien wird auf drei Ebenen vorgenommen: Die Vergleichbarkeit der Untersuchungsdesigns, die Vergleichbarkeit der Verfahrensauswahl und die Vergleichbarkeit der Stichproben. Es wird versucht, die vordergründig widersprüchlichen Ergebnisse der drei Studien auf diesen Ebenen zusammenzubringen, um sie sinnvoll interpretieren zu können.

Ad Vergleichbarkeit der Untersuchungsdesigns:

Bei der Versuchsplanung wurde dezidiert darauf geachtet, dass in allen drei Studien dasselbe Untersuchungsdesign verwendet wird, um eine Vergleichbarkeit zu ermöglichen. Den Studien ist gemein, dass jeweils ein Zwei-Gruppen-Versuchsdesign realisiert und die Variation der Vorgabereihenfolge auf identische Weise umgesetzt wurde. Die verwendeten Leistungstests und OPTs wurden jeweils geblockt variiert. Die Stichproben wiesen einen vergleichbaren Umfang auf, das Testsetting war ebenfalls sehr ähnlich. Alle drei Untersuchungen konnten in realen Auswahlprozessen im Unternehmenskontext durchgeführt werden. Das Vorgehen bei der Auswertung stimmte ebenfalls überein.

Obwohl das Untersuchungsdesign in hohem Maße vergleichbar war, wurden drei unterschiedliche Ergebnisse gefunden. Da auf dieser Betrachtungsebene eine so hohe Deckung der drei Studien besteht, erscheint es unwahrscheinlich, dass die unterschiedlichen Ergebnisse auf das Untersuchungsdesign zurückzuführen sind. Es müssen also andere Aspekte zur Erklärung der unterschiedlichen Ergebnisse herangezogen werden.

Ad Vergleichbarkeit der Verfahrensauswahl:

In allen drei Studien wurde zwar ein vergleichbares Untersuchungsdesign verwendet, bei der Verfahrensauswahl gab es im Detail jedoch bedeutende Unterschiede.

Die Verfahrensauswahl des Persönlichkeitsteils wurde in allen drei Studien nicht verändert, allerdings wurden konzeptionell ähnliche, jedoch nicht identische Verfahren im Leistungsteil eingesetzt. Die Matrizentests und die Intelligenztestbatterien wurden über alle drei Studien variiert. In Experiment 1 und 2 wurden unterschiedliche Matrizentests eingesetzt. Bei der Verwendung der SPM anstelle des AMTs in Experiment 2 trat kein Haupteffekt der Testreihenfolge auf, wohingegen bei Verwendung des AMTs in Experiment 1 ein Haupteffekt der Testreihenfolge auftrat, in Experiment 3 jedoch nicht. In Experiment 3 wurde erneut der AMT verwendet, um zu klären, ob dieser Test für die unterschiedlichen Ergebnisse in Experiment 1 und 2 verantwortlich ist. Es wurde also angenommen und auch von Khorramdel (2010) geäußert, dass diese beiden Matrizentests etwas mit Reihenfolgeeffekten zu tun

haben könnten. Die INSBAT und der IST 2000-R wurden ebenfalls zwischen den Studien variiert. Ein verzerrender Effekt der Variation der Intelligenztestbatterien ist zwar denkbar, aber aufgrund der in Abschnitt III.2.2.2. dargestellten Ähnlichkeit der vorgegebenen Subtests eher unwahrscheinlich. Im Hinblick auf die verwendeten Intelligenztestbatterien zeigt sich in Bezug zum Auftreten oder Ausbleiben von Reihenfolgeeffekten ebenfalls kein einheitliches Bild. Es lässt sich daraus abgeleitet vermuten, dass nicht die Verfahren selbst, sondern ein anderer Aspekt, der den hier vorgenommenen Verfahrensvariationen gemein ist, ausschlaggebend sein könnte.

AMT und SPM sind sich konzeptionell sehr nahe und unterscheiden sich vor allem durch ihren Schwierigkeitsgrad, der in der adaptiven und nicht-adaptiven Verfahrenskonstruktion begründet liegt. Die Items des SPM sind darüber hinaus allgemein leichter und adaptive Verfahren wie der AMT werden als subjektiv schwieriger erlebt (Abschnitt II.5.). Durch die Variation der INSBAT und des IST 2000-R wurde ebenfalls ein Austausch adaptiver und nicht-adaptiver Verfahren vorgenommen. Diese Überlegungen reduzieren die Unterschiede in der Verfahrensauswahl zwischen den Studien von einer Variation mehrerer Verfahren auf die Variation der Verfahrensgattung, der Itemschwierigkeit und möglicherweise unterschiedlicher Operationalisierungen psychologischer Konstrukte in den jeweiligen Verfahren. Diese Überlegungen deuten zwar neue Ansätze für Folgestudien an, können die inkonsistenten Ergebnisse bisher jedoch nicht weiterführend erklären. Es traten beispielsweise sowohl bei Verwendung adaptiver als auch nicht-adaptiver Tests Reihenfolgeeffekte auf.

Betrachtet man den Persönlichkeitsteil der Testbatterie, so wurden ausschließlich Objektive Persönlichkeitstests verwendet. In Experiment 1 wurden bei diesen Verfahren Reihenfolgeeffekte festgestellt. Allerdings zeigten sich die OPTs in den beiden Folgestudien (Experiment 2 und 3) robust gegenüber Reihenfolgeeffekten. Reihenfolgeeffekte erweisen sich also auch bei OPTs als nicht stabil.

Es wird ersichtlich, dass diese Überlegungen zur Verfahrensauswahl allein nicht ausreichen, um die inkonsistenten Ergebnisse auf einen Nenner zu bringen und nachvollziehbar erklären zu können. Andere - bisher nicht berücksichtigte - Aspekte der Verfahrensauswahl sowie Stichprobeneigenschaften könnten ebenfalls einen Einfluss haben.

Ad Vergleichbarkeit der Stichproben:

Die dritte Betrachtungsebene betrifft die Unterschiede und Gemeinsamkeiten der herangezogenen Stichproben. Anhand eines systematischen Vergleichs lassen sich Annahmen über mögliche Charakteristiken treffen, die für das Zustandekommen von Testreihenfolgeeffekten in Frage kommen.

Um die Stichproben aller drei Studien miteinander zu vergleichen, muss im Detail geklärt werden, wie sie sich unterscheiden und worin sie ähnlich charakterisiert sind. Während für die Stichproben der Experimente 1 und 2 Differenzen in der kognitiven Leistungsfähigkeit, dem Anspruchsniveau sowie der Belastbarkeit nachgewiesen wurden, können diesbezügliche Aussagen zur Stichprobe 3 nur eingeschränkt getätigt werden. Die Verwendung unterschiedlicher Matrizentests und Intelligenztestbatterien sowie unterschiedlicher Versionen des BACOs erschweren einen verlässlichen Vergleich der Werte über alle drei Studien. Es konnten jedoch Unterschiede in der Belastbarkeit, dem Anspruchs- und Leistungsniveau sowie der Flexibilität festgestellt werden. Die Betrachtung der unter Abschnitt III.2.4. genannten Eigenheiten der hier herangezogenen Stichprobe ermöglichen es dennoch einige Überlegungen und Vergleiche anzustellen.

Zu diskutieren sind das Bildungs- und Leistungsniveau, die Berufs- und Führungserfahrung, die Erfahrung mit Prüfungssituationen, die Testmotivation und die Belastbarkeit der Teilnehmer. In Experiment 3 wurde eine junge, motivierte mit Test- statt Führungserfahrung ausgestattete Stichprobe mit hohem Bildungs- und Leistungsniveau und ausgeglichenem Geschlechterverhältnis rekrutiert. Unter diesen Umständen traten keine Testreihenfolgeeffekte auf. Welche Unterschiede und Gemeinsamkeiten sind dann im Detail zu den anderen beiden Stichproben zu berichten?

Die Stichprobe aus Experiment 1 charakterisiert sich im Vergleich zur Stichprobe aus Experiment 2 wie bereits erwähnt durch ein hohes Intelligenz- und Anspruchsniveau, die Teilnehmer zeigten aber auch eine geringe Belastbarkeit in spezifischen Belastungssituationen. In dieser Kombination traten in Experiment 1 Testreihenfolgeeffekte auf. Bei der durch eine höhere Belastbarkeit ausgezeichneten Stichprobe in Experiment 2 trat kein Haupteffekt der Testreihenfolge auf. Für die Stichprobe aus Experiment 3 wurde aufgrund des Bildungsniveaus eine hohe kognitive Leistungsfähigkeit angenommen. Da keine Testreihenfolgeeffekte festgestellt

wurden, kann davon ausgegangen werden, dass die Testdauer keinen Einfluss auf die Leistungsfähigkeit der Teilnehmer hatte. Die Testleistung war über den Testverlauf in beiden Versuchsgruppen relativ konstant und der Bearbeitungszeitpunkt innerhalb der Testbatterie hatte keinen Einfluss auf die Leistung in den einzelnen Verfahren. Es kann also vor diesem Hintergrund von einer hohen Belastbarkeit in dieser Testsituation ausgegangen werden. Hinzu kommt, dass die Teilnehmer erst kürzlich ein Hochschulstudium beendet hatten und daher von einer ausgeprägten Erfahrung mit Prüfungs- oder Testsituationen ausgegangen werden kann (Abschnitt III.2.4.). Diese Umstände und das Ausbleiben von Testreihenfolgeeffekten in der vorliegenden Studie können im Sinne einer hohen Belastbarkeit der hier herangezogenen Stichprobe interpretiert werden. Ähnlich stellte sich die Situation in Experiment 2 dar, in der die Stichprobe ebenfalls eine höhere Belastbarkeit in den Kennwerten des BACOs aufwies als in Experiment 1, und Testreihenfolgeeffekte ausblieben. Khorramdel (2010) und Hambros (2002) vermuteten bereits die Anfälligkeit für Testreihenfolgeeffekte von Verfahren zur Erfassung von Belastbarkeit. Die Aufmerksamkeit wird somit auf Belastbarkeit als potentiell protektiven Faktor gegenüber Testreihenfolgeeffekten gelenkt.

Da das Versuchsdesign eine Konstante über alle drei Studien darstellt, verbleibt zur Erklärung der Ergebnisse nur das Zusammenwirken der beiden anderen Vergleichsebenen übrig. Die bisherigen Überlegungen und Erkenntnisse lassen vermuten, dass Reihenfolgeeffekte durch eine bisher nicht verlässlich verstandene und damit unkontrollierbare Kombination aus Stichprobeneigenschaften und Aspekten der Verfahrensauswahl zustande kommen, da beide Aspekte für sich genommen nicht zu einer Erklärung der inkonsistenten Ergebnisse aller drei Studien ausreichen.

Nach Betrachtung dieser drei Ebenen und unter Berücksichtigung der Erkenntnisse aus Experiment 1 und 2 kann für diese Studie ein Interpretationsansatz wie folgt formuliert werden: In der vorliegenden Studie traten keine Testreihenfolgeeffekte auf, da unter den durch das Versuchsdesign gegebenen Umständen die Stichprobe und die Testauswahl gut miteinander harmonisierten. Für die leistungsfähigen und hoch motivierten Testpersonen, denen aufgrund ihres erst kürzlich abgeschlossenen Studiums anspruchsvolle Testsituationen nicht fremd waren, war es irrelevant, in welcher Reihenfolge die Tests vorgegeben wurden. Sie waren in der Lage ihre

Leistungsfähigkeit, Konzentration und Motivation über einen langen Zeitraum auch bei hoher Anforderung aufrecht zu erhalten, da sie über eine hohe Belastbarkeit verfügten. Ob sich dieser Interpretationsversuch bewährt kann nur gezeigt werden, wenn er sich auf andere Ergebnisse anwenden lässt und in zukünftigen Studien erprobt wird.

Auf der dritten Stufe der Auseinandersetzung mit der vorliegenden Arbeit wird die im Theorieteil diskutierte Literatur umfassender berücksichtigt. Dort finden sich Anhaltspunkte für eine weiterführende Ausarbeitung des bisherigen Interpretationsansatzes, der Testreihenfolgeeffekte auf die Kombination von Stichprobeneigenschaften und Verfahrenstypen zurückführt.

Um den Interpretationsansatz weiter zu stützen, muss nachvollziehbar dargestellt werden, dass die Verfahren und somit die Testbatterie anspruchsvoll und die Stichprobe dieser Anforderung gewachsen war. Zhu und Tulskey (2000) konnten für fordernde Leistungstests bereits eine höhere Robustheit gegenüber Testreihenfolgeeffekte zeigen. Bei Baldinger (2006) und Khorramdel (2010) zeigten sich wenig fordernde Verfahren wie die AHA als anfällig für Reihenfolgeeffekte.

Für die in dieser Studie verwendeten Leistungstests kann angenommen werden, dass sie einen fordernden Charakter besitzen, wenn man die Auswirkungen adaptiver Testkonstruktion auf Leistung und Motivation der Testteilnehmer, wie sie in Abschnitt II.5. dargestellt wurden, betrachtet. Weniger anspruchsvolle Verfahren hingegen, können zu einer Unterforderung der Testteilnehmer führen, was negativen Einfluss auf die Testmotivation hat und auffälliges Antwortverhalten provoziert, wie es etwa bei Eiselt (1991) berichtet wurde.

Es stellt sich die berechtigte Frage, wie die im Vergleich zu den fordernden adaptiven Verfahren leichteren Aufgaben der OPTs, die in dieser Studie verwendet wurden, ebenfalls fordernd gewesen sein sollen. Vor allem die AHA zeichnen sich durch einfaches Itemmaterial aus. Man muss an dieser Stelle die Umstände der Testungen in allen drei Experimenten betrachten - die Testpersonen waren hochmotiviert. Bezieht man die Überlegungen zur Testmotivation aus Abschnitt II.7. mit ein, kann erklärt werden, warum auch den OPTs in diesen Studien ein fordernder Charakter zugeschrieben werden kann.

Mit dem BACO wurde ein OPT vorgegeben, der per Definition eine belastende Testsituation herstellt. Es wird darüber hinaus angenommen, dass die OPTs aufgrund motivationaler Aspekte in diesem Setting besonders ernst genommen wurden. Die Testpersonen können - der Konstruktion von OPTs entsprechend – aufgrund der hohen Augenscheinvalidität dieser Verfahren die Messintention zwar oft erraten, aber nicht gesichert feststellen, wie ihre Leistung verrechnet wird, woraus die geringere Verfälschbarkeit dieser Verfahren resultiert (Kubinger, 2009). Dies unterscheidet OPTs von den in der Regel leicht zu durchschauenden Persönlichkeitsfragebögen - es ist für die Testpersonen also nicht offensichtlich, in welchem Maße und auf welche Weise Persönlichkeitseigenschaften erfasst werden. Bei der Testung mit OPTs ist somit nicht einwandfrei zwischen Leistungs- und Persönlichkeitsvariablen zu differenzieren, wenn Unklarheit darüber herrscht, welche Kennwerte überhaupt herangezogen werden. Das Itemmaterial der hier verwendeten OPTs suggeriert in beabsichtigter Weise eine Leistungsanforderung - in bestimmten Subtests werden auch tatsächliche Leistungsvariablen im Sinne einer Persönlichkeitsdimension interpretiert werden (z. B. Exaktheit oder Leistungsniveau in den AHA). Es ist daher denkbar, dass die verwendeten OPTs den Testpersonen eher als eine Leistungstestung erschienen. So ist es etwa für die Testpersonen in der Gruppe LT → OPT nicht offensichtlich, dass nach Absolvierung des Leistungsteils keine Leistungskomponenten mehr erhoben werden, da diese in den hier verwendeten Subtests der OPTs weiterhin erforderlich sind. Hinzu kommt, dass die angenommene hohe Testmotivation dazu geführt haben könnte, dass die Teilnehmer auch diese Aufgaben genauso ernst genommen haben, wie die Leistungstests und auch die OPTs als dementsprechend fordernd erlebt haben.

In diesem speziellen Setting kann also angenommen werden, dass für OPTs die gleichen Regeln wie für fordernde Leistungstests gelten, das heißt, dass sie wenig anfällig für Reihenfolgeeffekte sind. Es kommt somit wieder die weiter oben erläuterte Robustheit fordernder Verfahren gegenüber Reihenfolgeeffekten zum Tragen. Diese Robustheit kann der hier dargelegten Argumentation folgend also ebenfalls für OPTs angenommen werden.

Obwohl in dieser Studie zwischen fordernden adaptiven Leistungstests und vergleichbar wenig anspruchsvollen OPTs gewechselt wurde, waren die Teilnehmer aus ihrer Perspektive vermeintlich stets mit fordernden Aufgaben konfrontiert. Die

Anforderung und Belastung der Testbatterie wurde dabei durch die Umstände der Auswahltestung verstärkt und nicht nur durch die Verfahren allein hergestellt. Es wird darüber hinaus vermutet, dass die hier herangezogene Stichprobe aufgrund ihrer ausgeprägten Test- und Prüfungserfahrung die vorgegeben Verfahren ebenfalls sehr ernst genommen und konzentriert bearbeitet hat, so dass ein Effekt einfacher und unterfordernder Aufgabentypen eventuell weniger zum Tragen kam.

Zusammengenommen kann vermutet werden, dass durch die hier vorgegebene Testbatterie aufgrund des Umfangs, der Auswahl-situation und durch die Vorgabe adaptiver Tests und spezieller OPTs eine hohe Anforderung an die Testpersonen gestellt wurde. Um dieser Belastung gewachsen zu sein, ist konsequenter Weise eine hohe Belastbarkeit erforderlich. Dies zu den in Abschnitt III.2.4. beschriebenen Stichprobencharakteristika in Bezug gesetzt, führt zu der Annahme, dass in einer durch ein hohes Leistungsniveau und ausgeprägte Testerfahrung charakterisierten Stichprobe bei der Bearbeitung einer anspruchsvollen Testbatterie keine Testreihenfolgeneffekte auftraten, da die Anforderung schlichtweg gut auf die Belastbarkeit der Stichprobe abgestimmt war. Die Testteilnehmer waren der Testbatterie gewachsen, die Verfahrensauswahl somit adäquat. So lässt sich eventuell auch erklären, dass bei Vorgabe der SPM und des IST 2000-R in einer kognitiv weniger leistungsfähigen Stichprobe ebenfalls kein Haupteffekt der Testreihenfolge gefunden wurde. Denn auch in Experiment 2 stimmten zumindest teilweise wieder die Leistungsfähigkeit und Belastbarkeit der Stichprobe und das Schwierigkeits- oder Anforderungsniveau der Verfahrensauswahl überein. Die Testpersonen der Experimente 1 und 2 verfügten vermutlich auch nicht über ein vergleichbares Maß an Testerfahrung aus Prüfungssituationen wie die Teilnehmer an Experiment 3.

Abschließend soll noch auf zwei weitere Aspekte der vorliegenden Arbeit eingegangen werden. Betrachtet man den Altersunterschied zwischen den Stichproben aus Experiment 3 im Vergleich zu den Experimenten 1 und 2 als kritisch, so lässt sich auf die Studie von Tulskey und Zhu (2000) verweisen, die keinen Zusammenhang von Testreihenfolgeneffekten und dem Alter der Testpersonen finden konnten. Hervorzuheben ist auch die nicht unerwartete Tatsache, dass in dieser Studie keine Ermüdungseffekte festgestellt wurden, obwohl die hier durchgeführten

Testungen zeitintensiv, fordernd und belastend waren. Ermüdungseffekte wurden von Khorramdel (2010) und Khorramdel und Frebort (2011) als mögliche Ursache für Testreihenfolgeeffekte vermutet. Der nicht signifikante Effekt der Testreihenfolgevariation in der vorliegenden Studie stützt eventuell auch die von Zhu und Tulsy (2000) formulierte und von Ackermann und Kanfer (2009) sowie Kahnemann (1973) gestützte Annahme, dass auch bei mehrstündigen Testungen keine Ermüdungseffekte auftreten, da die Testpersonen ihr Leistungsniveau aufrecht erhalten können, wenn es notwendig ist und sie dafür entsprechend motiviert sind.

Die ausführliche Diskussion der Ergebnisse hat deutlich gemacht, dass je nach Testsituation, Verfahrensauswahl und Stichprobe unterschiedliche Ergebnisse einer Testreihenfolgevariation zu beobachten sind. Bereits zu Beginn dieser Arbeit wurde das Anliegen formuliert, das bisherige Wissen um Testreihenfolgeeffekte anhand von Verfahrensgattungen und deren Kombination zu klassifizieren und zu systematisieren. Mit der folgenden tabellarischen Darstellung wird ein erster Versuch unternommen, dies umzusetzen. Dabei muss berücksichtigt werden, dass diese Darstellung lediglich der übersichtlichen Zusammenfassung der hier besprochenen Überlegungen dient und keine vollständig empirisch überprüften Ergebnisse darstellt. *Tabelle 12* enthält diese Zusammenstellung und leitet sich aus den vorangegangenen Überlegungen und den bereits in der Literatur verfügbaren Erkenntnissen ab. So sollen vor allem Denkanstöße für Folgestudien gegeben werden, es können aber auch Handlungsanweisungen für den Praktiker bei der Verfahrensauswahl in Personalauswahlprozessen entnommen werden.

Tabelle 12:

Vorschlag einer Systematisierung von Testreihenfolgeeffekten anhand von Verfahrensgattungen

<i>Instrument</i>	<i>Anfälligkeit für und Effekt von Testreihenfolgenvariationen</i>
Persönlichkeitsfragebögen	Hohe Anfälligkeit für eine Vielzahl inhaltlicher und formaler Reihenfolgeeffekte (siehe Abschnitt II.9.3.1. bis II.9.3.3.)
Kognitive Leistungstests	Geringe Anfälligkeit für Reihenfolgeeffekte in Personalauswahlprozessen bei stichprobenadäquat forderndem Aufgabenmaterial (siehe Abschnitt III.6.) Können bei inadäquater Anforderung Langeweile oder Erschöpfung hervorrufen, die darauf folgende Verfahren beeinflussen kann (siehe Abschnitt II.9.3.4.)
Objektive Persönlichkeitstests	Abhängig von der Art des Aufgabenmaterials Bei forderndem Aufgabenmaterial → Vergleichbare Robustheit wie kognitive Leistungstests Bei wenig forderndem Aufgabenmaterial → Vergleichbare Anfälligkeit wie Persönlichkeitsfragebögen

Obwohl in dieser Studie explizit nicht verwendet, wurden Persönlichkeitsfragebögen der Vollständigkeit halber bei der Darstellung berücksichtigt. Diese Verfahren erweisen sich anhand der in Abschnitt II.9.3.1. bis II.9.3.3. dargestellten Studien erwartungsgemäß als in hohem Maße anfällig für Testreihenfolgeeffekte.

Für kognitive Leistungstests und Objektive Persönlichkeitstests ist wichtig zu beachten, dass eine eindeutige Trennung zwischen freiwilligen und nicht-freiwilligen Stichproben in jedem Fall zu berücksichtigen ist. Die bisherigen Erkenntnisse belegen den bedeutsamen Unterschied zwischen diesen beiden Typen von Stichproben im Hinblick auf Testreihenfolgeeffekte. Leistungstests müssen getrennt zwischen adaptiven und nicht adaptiven Verfahren betrachtet werden. Aufgrund der besonderen Umstände einer Erhöhung der Stressbelastung und der gesteigerten Frustration durch adaptive Verfahren ist eine Verwendung im Leistungsbereich stets mit Vorsicht zu genießen. Es kann schnell zu einer Überforderung und Überlastung weniger leistungsfähiger Testpersonen kommen. Es kann jedoch vermutet werden, dass adaptive Verfahren für schlecht einschätzbare und vor allem auch heterogene

Stichproben ein probates Mittel darstellen, da sich ihr Anforderungsniveau zwar von Haus aus auf höherem Niveau bewegt, sich aber auch an der individuellen Leistungsfähigkeit der Testperson orientiert.

Hinsichtlich freiwilliger und nicht-freiwilliger Stichproben steht eine systematisch vergleichende Evaluation der Verfahrensgattung OPT noch aus. Eine Anfälligkeit ist hier aber zumindest theoretisch zu erwarten, insbesondere wenn die Verfahren wenig zeitgemäß sind und so einen verstärkenden negativen Effekt auf die Testmotivation haben.

Obwohl die vorliegende Studie zur weiteren Erforschung von Testreihenfolgeneffekte beigetragen hat und mit praktisch relevanten Schlussfolgerungen und Handlungsempfehlungen aufwarten kann, sind die Ergebnisse einigen Einschränkungen unterworfen. Zunächst ist wiederum auf die einschränkende Tatsache zu verweisen, dass die Stichprobe in wesentlichen Eigenschaften nicht wie vorgesehen mit Experiment 2 vergleichbar war. Wie auch in Experiment 1 und 2 wurde auf eine sehr spezifische Stichprobe zurückgegriffen. Auf die Unterschiede zwischen freiwilligen und nicht-freiwilligen Testteilnehmern wurde wiederholt hingewiesen, die Ergebnisse sind im strengen Sinne nur für nicht-freiwillige Stichproben generalisierbar. Da es sich bei der herangezogenen Stichprobe ausschließlich um Teilnehmer eines realen Auswahlverfahrens handelt, sind die Ergebnisse und Schlussfolgerung auch nur in diesen Bereichen zur Anwendung geeignet. Allerdings hebt dieser Umstand die praktische Relevanz der vorliegenden Arbeit für die Verbesserung von Auswahlprozessen im Unternehmenskontext besonders hervor.

Zukünftige Studien sollten dem Aspekt der Stichprobenspezifität weitere Beachtung schenken. Um eine weitere Generalisierung der Ergebnisse zu ermöglichen, sollte angestrebt werden, die Stichprobenbandbreite zu erweitern. Auch mehrere spezifische Stichproben ergeben irgendwann ein repräsentatives Gesamtbild. Darüber hinaus ist es im Sinne einer Vergleichbarkeit und praktisch relevanter Ergebnisse zu empfehlen, weiterhin auf nicht-freiwillige Stichproben zurückzugreifen.

Die Verfahrensauswahl und einige Stichprobencharakteristika wurden, an die Experimente 1 und 2 angelehnt, systematisch variiert, um die Aussagekraft der Ergebnisse der vorliegenden Studie durch den Vergleich mit ihren Vorgängern zu erhöhen. Es bleibt jedoch weiterhin nicht eindeutig geklärt, welche Verfahren konkret

für Reihenfolgeeffekte anfällig sind und welche Stichprobencharakteristika relevant sind, da diese nicht vollständig systematisch variiert werden konnten. Daher verbleibt die Interpretation der Ergebnisse auch auf Ebene der Verfahrenstypen. Einzelne Verfahren, die an der Entstehung von Reihenfolgeeffekten beteiligt sind, können auch über alle drei Studien nicht gesichert identifiziert werden. Es ist daher nicht auszuschließen, dass zwischen INSBAT und IST 2000-R sowie AMT und SPM weitere bedeutsame Unterschiede bestehen, die von dem hier herausgearbeiteten Interpretationsansatz nicht berücksichtigt werden. Komplexere Zusammenhänge zwischen den verwendeten Leistungstests und OPTs sind ebenfalls nicht auszuschließen. Bisher unbekannte oder nicht berücksichtigte Kontexteffekte könnten ebenfalls eine bedeutsame Rolle spielen.

In dieser Studie wurde eine blockweise Variation der vorgegebenen Verfahren realisiert, um eine Vergleichbarkeit zu den Experimenten 1 und 2 herzustellen und das Versuchsdesign nicht unnötig zu verkomplizieren. Für zukünftige Studien wäre es spannend, ob andere Formen der Vorgabereihenfolgenvariation, etwa eine Durchmischung der Persönlichkeits- und Leistungstests, unterschiedliche Auswirkungen haben. Es wäre somit erforderlich, weitere Studien durchzuführen, die nicht nur die Bandbreite der Stichproben, sondern auch die der Verfahren und deren Variation erweitern.

Besonders hervorzuheben ist, dass trotz aller interpretatorischen Bemühungen die Frage nach dem Einfluss der Matrizentests AMT und SPM nicht geklärt werden konnte. Weitere Studien mit tatsächlich vergleichbaren Stichproben und einer systematischen Variation dieser Verfahren sind hier zwingend erforderlich.

Somit steht noch vieles zur Disposition, gar nicht genauer erwähnt wird hierbei die mögliche Untersuchung einer Variation von Persönlichkeitsfragebögen in realen Auswahlprozessen, um der Verwendung dieser Verfahren in diesem Kontext nicht weiter Vorschub zu leisten.

Abschließend ist auf limitierende Umstände bei der Rekrutierung von Personen in realen Auswahlprozessen hinzuweisen, die sich auch bei der vorliegenden Studie nicht vollständig vermeiden ließen. Praktisch relevante Forschung sollte - wo möglich - in praktisch relevanten Kontexten durchgeführt werden. Dabei ist es notwendig, die Anliegen externer Kooperationspartner zu berücksichtigen, wodurch die wissenschaftliche Aussagekraft der Ergebnisse eingeschränkt werden kann. Diese

Anmerkung soll aber nicht abschrecken, es sollte lediglich im Vorfeld damit gerechnet werden, dass eventuell die Verfahrensauswahl oder die Stichprobenzusammensetzung nicht gänzlich der Kontrolle des Wissenschaftlers unterstehen. Die gesteigerte ökologischen Validität der Forschungsergebnisse wiegt diese Einschränkung jedoch in den meisten Fällen auf.

7. Zusammenfassung

In der vorliegenden Arbeit wurde die Fragestellung erörtert, welchen Einfluss eine Variation der Testreihenfolge in einer computerisierten Testbatterie auf die Ergebnisse in realen Auswahlprozessen hat. Dabei sollte vor allem an die zu unterschiedlichen Ergebnissen gekommenen Studien von Khorramdel und Frebort (2011) und Khorramdel (2010) angeknüpft werden und geklärt werden, ob die dort verwendeten Matrizentests für das Auftreten oder Ausbleiben von Testreihenfolgeeffekten verantwortlich sind. Es wurden bei den Überlegungen letztlich sowohl die verwendeten psychologisch diagnostischen Verfahren als auch Stichprobencharakteristika berücksichtigt.

Es wurde ein Experiment zum Einfluss von Testreihenfolgeeffekten in realen Auswahl Situationen durchgeführt. Eine Testbatterie, bestehend aus adaptiven Leistungstests und Objektiven Persönlichkeitstests, wurde 70 Bewerbern um ein Trainee-Programm der Österreichischen Bundesbahnen im Rahmen des Auswahlverfahrens vorgegeben. Die Teilnehmer wurden randomisiert in zwei Versuchsgruppen aufgeteilt, die jeweils eine unterschiedliche, nach Verfahrenstypen geblockte Testreihenfolge bearbeiteten. Sämtliche Kennwerte der verwendeten Verfahren wurden inferenzstatistisch auf Unterschiede zwischen den Versuchsgruppen überprüft. Die Analyse der erhobenen Daten mittels einer MANOVA erbrachte durchweg nicht signifikante Ergebnisse. Es konnte weder ein Haupteffekt der Testreihenfolgevariation ($p = 0,998$) noch in irgendeinem der erhobenen Kennwerten ein signifikanter Unterschied zwischen den Versuchsgruppen festgestellt werden ($0,981 > p > 0,151$). Dies bedeutet, dass in Bezug auf die Hypothesen 1, 2 und 3 (Abschnitt III.1.1.) jeweils die Alternativhypothesen zugunsten der Nullhypothesen abgelehnt wurden und dass in dieser Stichprobe ungeachtet der Bearbeitungsreihenfolge der Verfahrensblöcke von einem nicht statistisch bedeutsamen Einfluss der Testreihenfolge und des verwendeten Matrizentests (AMT) auszugehen ist.

Da die hier durchgeführte Studie hinsichtlich der Stichprobecharakteristika nicht wie beabsichtigt mit der von Khorramdel und Frebort (2011) verglichen werden konnte, wurden die Ergebnisse unter Hinzunahme beider Vorgängerstudien von Khorramdel (2010) und Khorramdel und Frebort (2011) diskutiert. Dabei wurden die Wechselwirkungen zwischen den verwendeten Verfahren und Testkonzepten sowie

den relevanten Stichprobeneigenschaften herausgearbeitet. Es wurde ein Interpretationsansatz vorgeschlagen, demzufolge die Wahl der Verfahren zwar das Auftreten von Testreihenfolgeeffekten begünstigen könne, dies aber letztlich maßgeblich von Eigenschaften der Stichprobe abhängig sei. So ist es gelungen, die vorliegenden nicht signifikanten Ergebnisse auf eine gute Übereinstimmung der Anforderung der Testbatterie und der Leistungsfähigkeit und Belastbarkeit der Stichprobe zurückzuführen. Dieser Interpretationsansatz wurde durch die im Theorieteil besprochenen Grundlagen gestützt und auch zur Erklärung der inkonsistenten Ergebnisse der Vorgängerstudien vorgeschlagen.

Die Verfahrensauswahl sollte bestmöglich auf Eigenheiten der Stichprobe abgestimmt werden, damit eine adäquat fordernde Testung gewährleistet werden kann. So kann möglicherweise das Auftreten von Testreihenfolgeeffekten vermieden werden. Dafür sollten die in dieser Studie diskutierten Eigenschaften wie Testerfahrung, Motivation, Belastbarkeit und das kognitive Leistungsniveau der Stichprobe berücksichtigt werden. Dementsprechend muss eine Verfahrensauswahl getroffen werden, die den Einfluss dieser Aspekte limitiert. Die Anforderung der Verfahren kann beispielsweise über die Verwendung adaptiver oder nicht-adaptiver Verfahren reguliert werden. Neben einer adäquaten Testauswahl sind selbstverständlich noch weitere bisher unberücksichtigte Aspekte denkbar, die das Auftreten oder Ausbleiben von Testreihenfolgeeffekten beeinflussen können.

Bei der Personalauswahl im Unternehmenskontext ist vorrangig mit nicht-freiwilligen Stichproben zu rechnen, für die eine hohe Testmotivation angenommen werden kann. Es sollte bei der weiterführenden Erforschung von Testreihenfolgeeffekten ebenfalls möglichst auf nicht-freiwillige Stichproben zurückgegriffen werden. Motivationale Aspekte sind in diesen Stichproben weniger problematisch. Eine geringe Testmotivation kann bei inadäquater Verfahrensauswahl das Auftreten von Reihenfolgeeffekten begünstigen.

Die bisher ausgegebene Empfehlung, in Auswahl Situationen aufgrund der ungewissen Auswirkungen von einer Testreihenfolgevariation abzuraten, kann anhand der hier gewonnenen Erkenntnisse kaum aufgelockert werden. Das inkonsistente Auftreten von Testreihenfolgeeffekten mündet weiterhin in der Empfehlung in Auswahlprozessen die Testreihenfolge konstant zu halten, um einen die Fairness einschränkenden Effekt zu vermeiden.

Diese Arbeit zeigt darüber hinaus, dass eine weitere Erforschung von Testreihenfolgeeffekten konsequenterweise notwendig ist, da das Auftreten und Ausbleiben dieser Effekte nicht abschließend geklärt werden konnte. Vorsicht ist vor allem immer dann geboten, wenn unklar ist, wie die Stichprobe hinsichtlich der für Testreihenfolgeeffekte relevanten Eigenschaften tatsächlich charakterisiert ist oder wenn das Anforderungsniveau der Testbatterie schwierig abzuschätzen ist. Anhand der in dieser Arbeit vorgeschlagenen Systematisierung bisheriger Forschungsergebnisse können psychologisch diagnostische Verfahren besser hinsichtlich ihrer Anfälligkeit für Testreihenfolgeeffekte eingeschätzt werden. Diese Zusammenstellung erleichtert den Überblick über die bisherigen Erkenntnisse zu Testreihenfolgeeffekten und erlaubt es, Handlungsanweisungen sowohl für die weitere Forschung als auch für die Verfahrensauswahl in der Praxis abzuleiten.

IV. Anhänge

1. Rückmeldebogen



Potentialanalyse Ergebnisrückmeldung



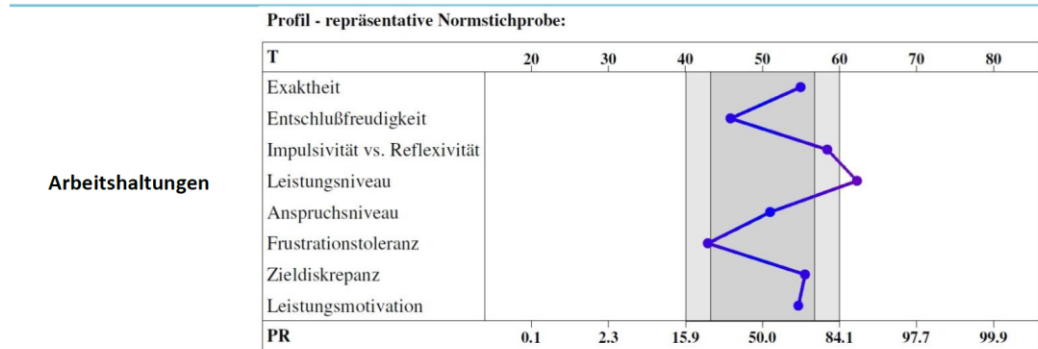
Zusammenfassung - Überblick

Vorname / Nachname	██████████																
Alter	25 Jahre																
Angestrebte Stelle	Trainee																
Erläuterung Prozentrang (PR)	Dies ist der Prozentwert der Personen der Vergleichsstichprobe mit gleicher oder niedrigerer Ausprägung in der betreffenden Dimension. Der Durchschnittsbereich ist im unten stehenden Diagramm zwischen PR 33 und 66 markiert, Werte darunter gelten als unter-durchschnittlich und Werte darüber als überdurchschnittlich																
Übersicht	<table border="1"> <thead> <tr> <th>Dimension</th> <th>Prozentrang</th> </tr> </thead> <tbody> <tr> <td>Allgemeine Intelligenz</td> <td>96</td> </tr> <tr> <td>Fluide Intelligenz</td> <td>98</td> </tr> <tr> <td>Quantitatives Denken</td> <td>97</td> </tr> <tr> <td>Kurzzeitgedächtnis</td> <td>100</td> </tr> <tr> <td>Impulsivität vs. Reflexivität</td> <td>80</td> </tr> <tr> <td>Anspruchsniveau</td> <td>54</td> </tr> <tr> <td>Leistungsniveau</td> <td>89</td> </tr> </tbody> </table>	Dimension	Prozentrang	Allgemeine Intelligenz	96	Fluide Intelligenz	98	Quantitatives Denken	97	Kurzzeitgedächtnis	100	Impulsivität vs. Reflexivität	80	Anspruchsniveau	54	Leistungsniveau	89
Dimension	Prozentrang																
Allgemeine Intelligenz	96																
Fluide Intelligenz	98																
Quantitatives Denken	97																
Kurzzeitgedächtnis	100																
Impulsivität vs. Reflexivität	80																
Anspruchsniveau	54																
Leistungsniveau	89																
Zusammenfassung	Die Testperson zeigte unter Belastung durch Aufgabenkollision ein ausgewogenes Verhältnis bei der Bearbeitung von Haupt- und Nebenaufgaben. Insgesamt arbeitete die Testperson in Belastungssituationen überdurchschnittlich effizient. Bei Belastung durch Verhinderung des planmäßigen Vorgehens gelang es der Testperson trotz laufender Unterbrechungen ein vorgegebenes Ziel zu erreichen. Hierbei setzte sie viele Strategien ein, die zu einer schnellen Zielerreichung führen und zeigte außerdem durchschnittlich viele nicht zielführende Verhaltensweisen, die sich in der Wiederholung nicht erfolgreicher Lösungsversuche äußern. Bei der Tagesplanung ging die Testperson strukturiert jedoch wenig effizient vor und erledigte einige der vorgegebenen und wenige der vorgenommenen Aktivitäten. Insgesamt handelte die Testperson durchschnittlich zielorientiert, sehr vermittelnd und war in ihren Handlungen in unauffälligem Maße ablenkbar.																
Beobachtungen	Keine besonderen Vorkommnisse.																
Für die sachliche Richtigkeit Leonard Schünemann – Mag. Simon Lehner																	

Vorname / Nachname								
Erläuterung Prozentrang (PR)		Dies ist der Prozentwert der Personen der Vergleichsstichprobe mit gleicher oder niedrigerer Ausprägung. Werte zwischen 25 und 75 gelten als durchschnittlich, Werte darunter als unter-durchschnittlich und Werte darüber als überdurchschnittlich						
Legende	Unauffällig	Stärke	Schwäche	Neutral				
Profil - Repräsentative Normstichprobe (ab 14 Jahre, Bildungsgrad 4-5):								
INSBAT	T	20	30	40	50	60	70	80
	Allgemeine Intelligenz							
	Fluide Intelligenz							
	Quantitatives Denken							
	Kurzzeitgedächtnis							
	Numerisch-induktives Denken							
	Verbal-deduktives Denken							
	Wortbedeutung							
	Verbales Kurzzeitgedächtnis							
	Visuelles Kurzzeitgedächtnis							
	Arithmetische Kompetenz							
	Arithmetische Flexibilität							
PR		0.1	2.3	15.9	50.0	84.1	97.7	99.9
Kennwert		PR						
Allgemeine Intelligenz	96	Diese Variable ermöglicht eine Globalbeurteilung der intellektuellen Leistungsfähigkeit einer Person. Personen mit einer hohen Ausprägung in dieser Variable verfügen allgemein über ein hohes Maß an intellektueller Leistungsfähigkeit.						
Fluide Intelligenz	98	Diese Variable erlaubt eine Globalbeurteilung der Fähigkeit, allgemeine Gesetzmäßigkeiten zu erkennen, Implikationen zu verstehen und logische Schlussfolgerungen anhand numerischen, figural-bildlichen oder verbalen Aufgabenmaterials zu ziehen.						
Numerisch-induktives Denken	86	Diese Variable misst logisch-induktives Denken im Umgang mit numerischem Itemmaterial. Sie erfasst die Fähigkeit, logische Gesetzmäßigkeiten zu erkennen und zu abstrahieren. Personen mit einer hohen Ausprägung erkennen rasch Regelmäßigkeiten in Arbeitsabläufen und können von ihren bisherigen Erfahrungen abstrahieren.						
Verbal-deduktives Denken	98	Verbal-deduktives Denken ist die Fähigkeit zum formal logischen Umgang mit verbalen Aussagen. Personen mit einer hohen Ausprägung sind fähig, den Inhalt von verbalen Aussagen - unabhängig von ihrem eigenen Erfahrungshintergrund - zu erfassen, zu verknüpfen und daraus der Logik entsprechende Schlüsse zu ziehen.						

Wortbedeutung	69	Diese Variable erfasst den Umfang des passiven Wortschatzes einer Person.
Kurzzeitgedächtnis	100	Diese Variable dient der Globalbeurteilung der Fähigkeit, Informationen visueller und verbaler Art kurzfristig behalten und reproduzieren zu können.
Verbales Kurzzeitgedächtnis	99	Diese Variable misst die Fähigkeit, verbale Informationen kurzfristig zu speichern und richtig wieder abzurufen. Ein hoher Wert bedeutet, dass die Person einen größeren Umfang an verbalen Informationen kurzfristig behalten und anschließend akkurat reproduzieren kann.
Visuelles Kurzzeitgedächtnis	97	Diese Variable misst die Fähigkeit, visuelle Informationen kurzfristig zu speichern und richtig wieder abzurufen. Ein hoher Wert bedeutet, dass die Person einen größeren Umfang an visuellen Informationen kurzfristig behalten und anschließend reproduzieren kann.
Quantitatives Denken	97	Diese Variable dient der Globalbeurteilung des Verständnisses grundlegender mathematischer Prinzipien und Fertigkeiten und der Fähigkeit, diese zur Lösung praktischer Problemstellungen anwenden zu können. Personen mit einer hohen Ausprägung in dieser Fähigkeitsdimension verfügen somit über ein ausgereiftes Verständnis des dekadischen Aufbaus von Zahlen und der Anwendbarkeit und Wirkungsweise von Rechenoperationen. Sie sind zudem dazu in der Lage, mit Hilfe dieses Wissens Situationen mathematisch zu modellieren.
Arithmetische Kompetenz	95	Fähigkeit zum schnellen und automatisierten Umgang mit Zahlen und einfachen bis komplexen arithmetischen Rechenoperationen. Eine hohe Ausprägung kann als Indiz für ein umfangreiches arithmetisches Wissen, sowie die Verfügbarkeit effizienter Kopfrechenstrategien, angesehen werden.
Arithmetische Flexibilität	60	Diese Variable erfasst das Verständnis der Anwendbarkeit der Grundrechenarten, sowie die kognitive Flexibilität im Umgang mit numerischem Aufgabenmaterial. Ein hoher Wert zeigt an, dass die Person über ein gut strukturiertes und umfangreiches Wissen über die funktionalen Möglichkeiten der Grundrechenarten verfügt und dieses auch flexibel zur Lösung formal-mathematischer Problemsituationen anwenden kann.

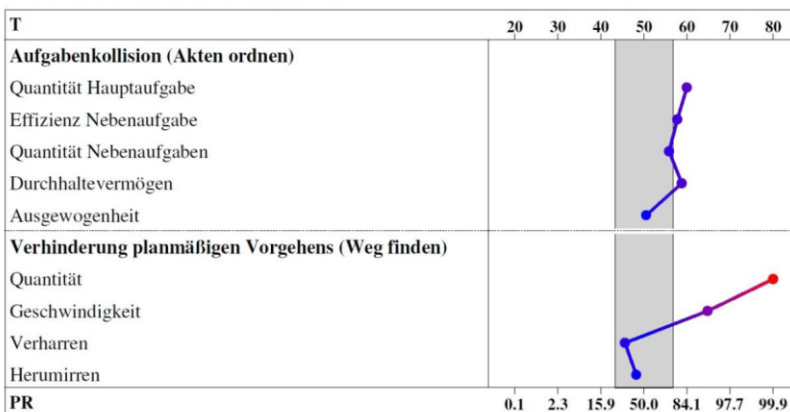
AMT		
Allgemeine Intelligenz	94	Dies ist die Hauptvariable, mit der die Fähigkeit zum sprachfreien induktiv logischen Schließen erfasst wird. Eine Person mit hoher Ausprägung in dieser Variable besitzt in besonderem Maß die Fähigkeit, Gesetzmäßigkeiten bzw. Regelmäßigkeiten zu erkennen und die daraus abgeleiteten Regeln anzuwenden. Personen mit einem überdurchschnittlichen Prozentrang können Gesetzmäßigkeiten aus ihrer Lernerfahrung abstrahieren und daraus Konsequenzen für zukünftiges Verhalten ableiten.



Kennwert	PR	
Exaktheit	69	Je höher dieser Wert, desto größer ist die Exaktheit .
Entschlußfreudigkeit	34	Je höher dieser Wert, desto größer ist die Entschlußfreudigkeit.
Impulsivität vs. Reflexivität	80	Diese Variable ist so zu interpretieren, dass ein geringer Wert Ausdruck von Impulsivität ist, bzw. ein hoher Wert Ausdruck von Reflexivität ist.
Leistungsniveau	89	Ein hoher Wert bedeutet, dass der Proband in der zweiten Etappe, die auch als Niveauanker dient, viele Symbole bearbeitet hat. Dies kann ein Hinweis auf gute Konzentrationsleistung bei Kodieraufgaben und/oder auf guten Umgang mit der Maus sein.
Anspruchsniveau	54	Mit dieser Variable wird festgestellt, ob der Proband zu realistischen oder unrealistischen Zielsetzungen neigt. Unterdurchschnittliche sowie überdurchschnittliche Prozentränge lassen auf unrealistische Zielsetzungen schließen.

Frustrationstoleranz	24	Diese Variable drückt aus, inwieweit sich der Proband von wiederholten negativen Rückmeldungen beeinflussen lässt. Erfolgt im Verlauf des Tests eine Korrektur der Prognosen nach unten, so resultiert ein unterdurchschnittlicher Prozentrang, der als geringe Frustrationstoleranz interpretiert werden kann.
Zieldiskrepanz	71	Die Zieldiskrepanz gibt an, ob die Leistung und die Prognose in einem vernünftigen Verhältnis zueinander stehen. Sehr kleine Werte sind ein Zeichen für große Übereinstimmung, sehr große für starke Diskrepanz zwischen Leistung und Prognose. Die Richtung der Abweichung wird dabei nicht berücksichtigt.
Leistungsmotivation	68	Diese Variable beinhaltet die „Anzahl richtiger Antworten“. Je größer der Wert ist, desto größer ist auch die Leistungsmotivation

Profil - Normstichprobe (Bildungsgrad 4-5):



Kennwert	PR	
<u>Aufgabenkollision</u>		
Quantität Hauptaufgabe	84	Personen mit einem über- (unter-) durchschnittlichen Testwert in dieser Variable erfüllen in einer Situation unter Aufgabenkollision eine als Hauptaufgabe deklarierte Tätigkeit über- (unter-) durchschnittlich gut.
Effizienz Nebenaufgabe	78	Personen mit einem über- (unter-) durchschnittlichen Testwert in dieser Variable zeigen unter Aufgabenkollision eine über- (unter-) durchschnittlich effiziente Strategie, mit auftretenden Unterbrechungen umzugehen.
Quantität Nebenaufgabe	72	Personen mit einem über- (unter-) durchschnittlichen Testwert in dieser Variable erfüllen in einer Situation unter Aufgabenkollision eine als Nebenaufgabe deklarierte Tätigkeit über- (unter-) durchschnittlich gut.
Durchhaltevermögen	81	Personen mit einem über- (unter-) durchschnittlichen Testwert in dieser Variable zeigen nach einer bestimmten Belastungsdauer eine über-(unter-) durchschnittliche Leistungsfähigkeit hinsichtlich der Hauptaufgabe.
Ausgewogenheit	52	Ein mittlerer Wert zeigt, dass Testpersonen beiden Aufgabenaspekten gerecht werden. Abweichende Ergebnisse deuten darauf hin, dass je ein Ziel zugunsten eines anderen höher bewertet wurde.
<u>Verhinderung des planmäßigen Vorgehens</u>		
Quantität	100	Personen mit einem über- (unter-) durchschnittlichen Testwert in dieser Variable gelingt es über- (unter-) durchschnittlich leicht, in einer Situation, in welcher geplantes Vorgehen laufend unterbrochen wird, ein vorgegebenes Ziel zu erreichen.

Geschwindigkeit	93	Personen mit einem über- (unter-) durchschnittlichen Testwert in dieser Variable zeigen bei Verhinderung des planmäßigen Vorgehens Strategien, welche zu einer über- (unter-) durchschnittlich schnellen Zielerreichung führen.
Verharren	33	Personen mit einem über- (unter-) durchschnittlichen Testwert in dieser Variable weisen unter Verhinderung des planmäßigen Vorgehens über-(unter-) durchschnittlich häufig Zeitspannen ohne aufgabenbezogene Aktion auf.
Herumirren	43	Personen mit einem über- (unter-) durchschnittlichen Testwert in dieser Variable zeigen bei Unterbrechungen bzw. Verhinderung über-(unter-) durchschnittlich häufig ein nicht zielführendes Verhalten, welches sich in einer Wiederholung nicht erfolgreicher Lösungsversuche zeigt.
Zielerreichung	Ja	Personen, welche das Ziel (nicht) erreicht haben, haben (nicht) alle Hindernisse im Verlauf bewältigt und vorher abgebrochen.

ILICA

Kennwert	PR/kategorial	
Erfüllung der gestellten Aufgaben	Nicht erfüllt	Diese Variable unterscheidet zwischen Personen, die alle 5 Ihnen gestellten Aufgaben erfüllt bzw. nicht erfüllt haben.
Strategie	Mit Tagesplanung	Die Testperson hat die Auswahl den Tag mithilfe eines Planers zu strukturieren oder spontan zu handeln.
Aktivitäten gesamt	46	In dieser Variable resultiert ein umso höherer Prozentrang, je mehr Aktivitäten ausgeführt wurden.
Geplante Aktivitäten	21	Diese Variablen schlüsseln die Gesamtaktivitäten in geplante, geplante und ausgeführte und ungeplante Aktivitäten auf. Hohe Prozenzränge resultieren, je mehr Aktivitäten dieser Art vorgenommen wurden.
Geplante und ausgeführte Aktivitäten	43	
Ungeplante Aktivitäten	34	
Orientierung	Kalender 21 Uhr 97 Telefon 55	Diese Variable gibt an, anhand welcher Hilfsmittel sich die Testperson orientiert. Hohe Prozenzränge resultieren, je öfter diese Option gewählt wurde.
Anteil spezifischer Aktivitäten	Urlaub 36 Freizeit 37 Alltag 82 Sozial 72	Diese Variable schlüsselt die Gesamtaktivitäten in die Kategorien Urlaub, Freizeit, Alltag und Sozial auf. Hohe Prozenzränge bedeuten, dass die Testperson im Vergleich zu anderen prozentuell öfter Aktivitäten der fraglichen Art gesetzt hat.
Resultierender Typ	Leicht Ablenkbar 53 Vermittelnd 79 Zielorientiert 33	Diese Variable differenziert zwischen den Typen „Vermittelnd“ – „Leicht Ablenkbar“ und „Zielorientiert“. Ein hoher Prozenzrang bedeutet, dass die Testperson im Vergleich zu anderen prozentuell öfter in betreffender Richtung reagiert hat.

2. Syntax für z-Transformation in SPSS

RECODE

Kennwert in Prozentrang

(0=-3.719) (1=-2.326) (2=-2.054) (3=-1.881) (4=-1.751) (5=-1.645) (6=-1.555) (7=-1.476) (8=-1.405) (9=-1.341) (10=-1.282) (11=-1.227) (12=-1.175) (13=-1.126) (14=-1.080) (15=-1.036) (16=-0.994) (17=-0.954) (18=-0.915) (19=-0.878) (20=-0.842) (21=-0.806) (22=-0.772) (23=-0.739) (24=-0.706) (25=-0.674) (26=-0.643) (27=-0.613) (28=-0.583) (29=-0.553) (30=-0.524) (31=-0.496) (32=-0.468) (33=-0.440) (34=-0.412) (35=-0.385) (36=-0.358) (37=-0.332) (38=-0.305) (39=-0.279) (40=-0.253) (41=-0.228) (42=-0.202) (43=-0.176) (44=-0.151) (45=-0.126) (46=-0.100) (47=-0.075) (48=-0.050) (49=-0.025) (50=0) (51=0.025) (52=0.050) (53=0.075) (54=0.100) (55=0.126) (56=0.151) (57=0.176) (58=0.202) (59=0.228) (60=0.253) (61=0.279) (62=0.305) (63=0.332) (64=0.358) (65=0.385) (66=0.412) (67=0.440) (68=0.468) (69=0.496) (70=0.524) (71=0.553) (72=0.583) (73=0.613) (74=0.643) (75=0.674) (76=0.706) (77=0.739) (78=0.772) (79=0.806) (80=0.842) (81=0.878) (82=0.915) (83=0.954) (84=0.994) (85=1.036) (86=1.080) (87=1.126) (88=1.175) (89=1.227) (90=1.282) (91=1.341) (92=1.405) (93=1.476) (94=1.555) (95=1.645) (96=1.751) (97=1.881) (98=2.054) (99=2.326) (100=3.719)

INTO *Kennwert in z-Wert*.

EXECUTE .

3. Diskriminanzanalyse

Tabelle 13:

Eigenwerte der Diskriminanzanalyse – Da nur der Faktor Testreihenfolge existiert, kann nur eine Diskriminanzfunktion getestet werden.

Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	,260	100,0	100,0	,454

Tabelle 14:

Signifikanzprüfung der ersten Diskriminanzfunktion – $p = 0,986$ ist nicht signifikant

Wilks Λ

Test der Funktion	Wilks Λ	X^2	df	Signifikanz
1	,793	12,722	26	,986

Tabelle 15:

Koeffizienten der ersten Diskriminanzfunktion

Standardisierte kanonische Diskriminanzfunktionskoeffizienten

Exaktheit	,583	Ausgewogenheit	-,314
Entschlussfreudigkeit	,111	Effizienz Nebenaufgabe	,697
Impulsivität vs. Reflexivität	-,066	Durchhaltevermögen	,065
Leistungsniveau	,614	Verbal Deduktives Denken	-,697
Anspruchsniveau	-,084	Arithmetische Kompetenz	,645
Frustrationstoleranz	,537	Verbales Kurzzeitgedächtnis	,035
Zieldiskrepanz	,215	Visuelles Kurzzeitgedächtnis	-,119
Allgemeine Intelligenz	-,069	Numerisch Induktives Denken	,049
Quantität	-,510	Arithmetische Flexibilität	-,159
Verharren	,150	Wortbedeutung	,046
Herumirren	-,345	Flexibel - vermittelnd	8,005
Quantität Hauptaufgabe	,010	Impulsiv - leicht abzulenken	7,370
Quantität Nebenaufgaben	,099	Reflexiv - zielorientiert	8,204

Die mittlerweile veraltete Version BaCO-D wurde in Experiment 1 und 2 verwendet. In Experiment 3 wurde die Nachfolgeversion BACO verwendet

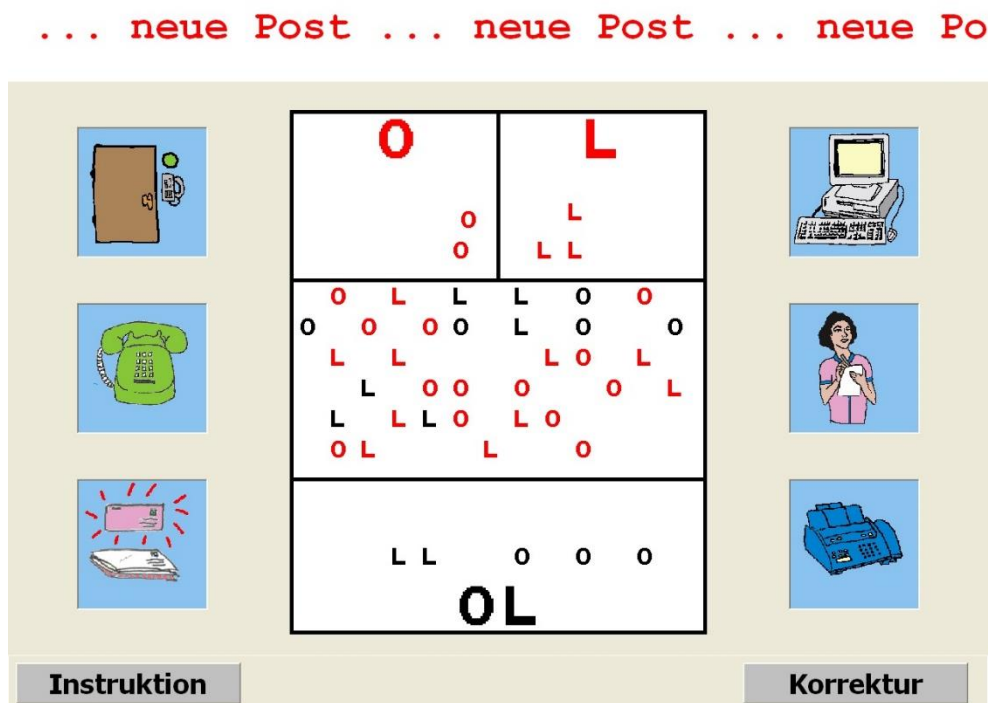


Abbildung 3: Subtest Belastbarkeit bei Aufgabenkollision im BaCO-D in der veralteten Version aus Experiment 1 und 2

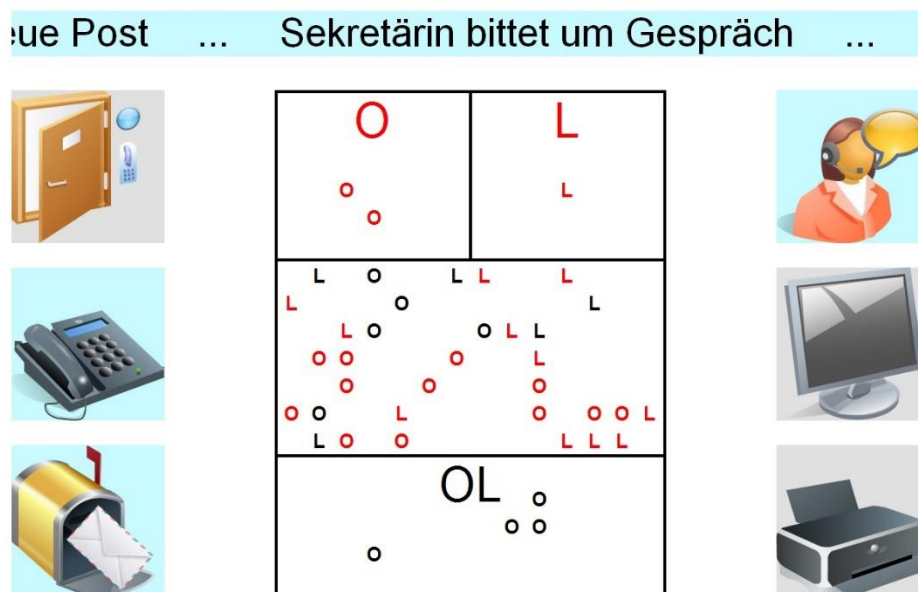


Abbildung 4: Subtest Belastbarkeit bei Aufgabenkollision im BACO in der neuen Version aus Experiment 3

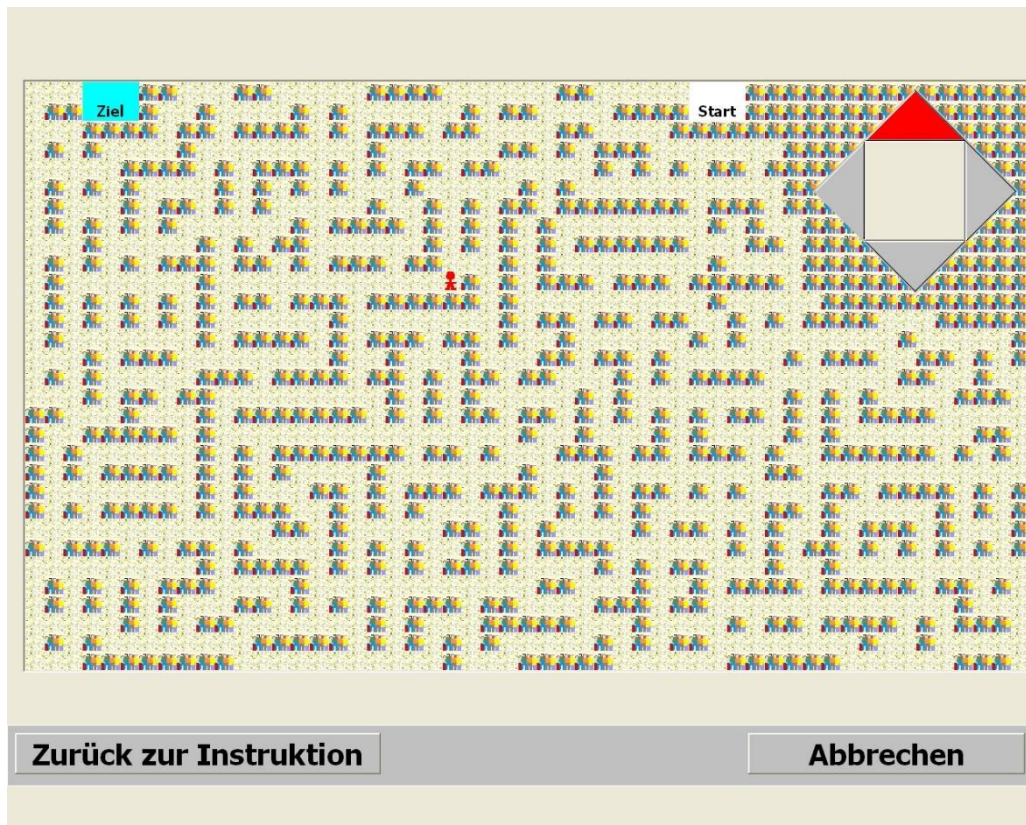


Abbildung 5: Subtest Belastbarkeit bei Verhinderung des planmäßigen Vorgehens im BaCO-D in der veralteten Version aus Experiment 1 und 2

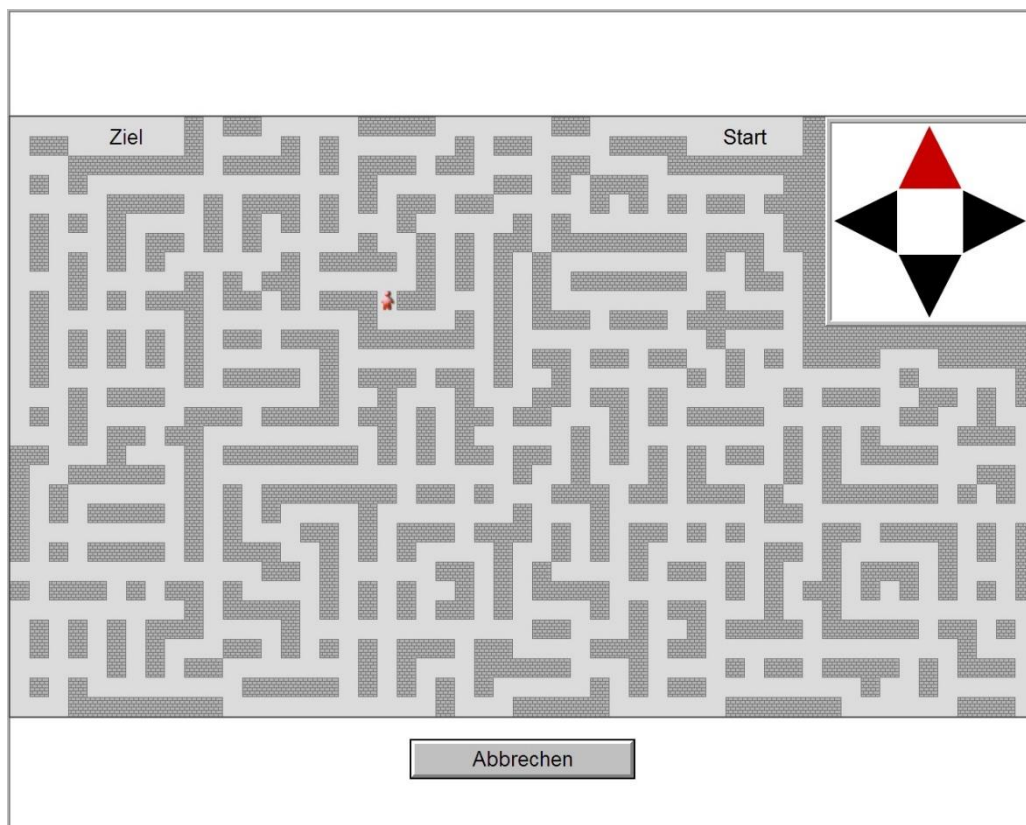


Abbildung 6: Subtest Belastbarkeit bei Verhinderung des planmäßigen Vorgehens im BaCO in der neuen Version aus Experiment 3

5. Tabellenverzeichnis

<i>Tabelle 1:</i>	22
<i>Tabelle 2:</i>	60
<i>Tabelle 3:</i>	73
<i>Tabelle 4:</i>	83
<i>Tabelle 5:</i>	84
<i>Tabelle 6:</i>	88
<i>Tabelle 7:</i>	91
<i>Tabelle 8:</i>	93
<i>Tabelle 9:</i>	94
<i>Tabelle 10:</i>	95
<i>Tabelle 11:</i>	96
<i>Tabelle 12:</i>	109
<i>Tabelle 13:</i>	123
<i>Tabelle 14:</i>	124
<i>Tabelle 15:</i>	124

6. Abbildungs- und Formelverzeichnis

<i>Abbildung 1:</i>	55
<i>Abbildung 2:</i>	86
<i>Abbildung 3:</i>	125
<i>Abbildung 4:</i>	125
<i>Abbildung 5:</i>	126
<i>Abbildung 6:</i>	126
<i>Formel 1:</i>	77
<i>Formel 2:</i>	77

7. Literaturverzeichnis

- Ackerman, P. L. (2000). Domain-specific knowledge as the “dark matter” of adult intelligence: Gf/gc, personality and interest correlates. *Journal of Gerontology: Psychological Sciences*, 55, 69-84.
- Ackerman, P. L. & Kanfer, R. (2009). Test Length and Cognitive Fatigue: An Empirical Examination of Effects on Performance and Test-Taker Reactions. *Journal of Experimental Psychology: Applied*, 15, 163-181.
- Anger, H. (1969). Befragung und Erhebung. In C. F. Graumann. (Hrsg.), *Handbuch der Psychologie*, 7, (S.567-618). Göttingen: Hogrefe.
- Arendasy, M., Hornke, L. F., Sommer, M., Häusler, J., Wagner-Menghin, Gittler, Bogner & Wenzl, M. (2005). *Manual Intelligenz-Struktur-Batterie (INSBAT)*. Mödling: SCHUHFRIED GmbH.
- Asseburg, R. (2011). *Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests*. Unveröffentlichte Dissertation, Universität Kiel.
- Baehr, M. E. (1953). A simplified procedure for the measurement of employee attitudes. *The Journal of Applied Psychology*, 37,163-167.
- Baldinger, D. (2006). *Der Einfluss sozial erwünschten Verhaltens auf das Ergebnis Objektiver Persönlichkeitstests*. Unveröffentlichte Diplomarbeit, Universität Wien.
- Becker, P. (1989). *Der Trierer Persönlichkeitsfragebogen (TPF)*. Göttingen: Hogrefe.
- Betz, N. E. (1975). New types of information and psychological implications. In D. J. Weiss (Hrsg.), *Computerized adaptive trait measurement: Problems and Prospects* (Forschungsbericht 75-5, S. 32–43). Minneapolis: University of Minnesota.
- Betz, N. E. & Weiss, D. J. (1976). Psychological effect of immediate knowledge of results and adaptive ability testing (Forschungsbericht 76-4). Minneapolis: University of Minnesota.
- Brandon, A. D. & Chavez, E. L. (1985). Order and delay effects on neuropsychological test presentation: The Halstead Category and Wisconsin Card Sorting Tests. *The International Journal of Clinical Neuropsychology*, 7, 152-153

- Cassel, R. H., Johnson, A. P. & Burns, W. H. (1962). The order of tests in the battery. *Journal of Clinical Psychology*, 18, 464-465.
- Cattell, R. B. (1957). *Personality and motivation: Structure and measurement*. New York: World Book
- Cattell, R. B. (1958). What is "objective" in "objective" personality tests? *Journal of Counseling Psychology*, 5, 285–289.
- Collani, G. von & Schyns, B. (2002). Generelle Selbstwirksamkeitserwartung. In A. Glöckner-Rist (Hrsg.), *ZUMA-Informationssystem. Elektronisches Handbuch sozialwissenschaftlicher Erhebungsinstrumente. Version 6.00*. Mannheim: Zentrum für Umfragen, Methoden und Analysen.
- Dalbert, C., Strack, K., Boesken, W. H. & Schwenkmezger, P. (1991). Beeinflusst die Auseinandersetzung mit Bewältigungsverhalten das Angsterleben? *Zeitschrift für klinische Psychologie*, 20, 65-74.
- Davis, D. R. (1946). The disorganization of behaviour in fatigue. *Journal of Neurology, Neurosurgery and Psychiatry*, 9, 23–29.
- DIN Deutsches Institut für Normung e.V. (2002). *Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen: DIN 33430*. Berlin: Beuth.
- Dodge, R. (1917). The laws of relative fatigue. *Psychological Review*, 24, 89-113
- Dörner, N. (2009). *Über den Einfluss der inhaltshomogenen Itemblockbildung auf die Dimensionalität eines Persönlichkeitsfragebogens*. Unveröffentlichte Diplomarbeit, Universität Wien.
- Eiselt, W. (1991). *Der Einfluss subjektiver Daten und Testreihenfolge auf die Leistung in einer computergestützten Testbatterie*. Forschungsbericht des Schiffahrtmedizinisches Instituts der deutschen Marine, Kiel, 5-54.
- Eisenhauer, E. (2008). *Effekte bei der Veränderung der Itempositionen anhand des Anstrengungsvermeidungstest und B5PO*. Unveröffentlichte Diplomarbeit, Universität Wien.
- Eysenck, H.-J., Wilson, C. D. & Jackson, C. J. (1998). *Eysenck Personality Profiler (EPP-D)*. Frankfurt am Main: Swets.

- Fahrenberg, J., Hampel, R. & Selg, H. (1984). *Das Freiburger Persönlichkeitsinventar - Revidierte Fassung (FPI-R)*. Göttingen: Hogrefe.
- Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149-1160.
- Faust, D. S. & Hollingsworth, J. O. (1991). Concurrent validation of the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R) with two criteria of cognitive abilities. *Journal of Psychoeducational Assessment*, 9, 224-229.
- Fazio, R. H., Powell, M. C. & Herr, P. M. (1983). Toward a process model of the attitude behavior relation: Accessing one's attitude upon mere observation of the attitude object. *Journal of Personality and Social Psychology*, 44, 723-735.
- Földényi, M., Tagwerker-Neuenschwander, F., Giovanoli, A., Schallberger, U. & Steinhausen, H.-C. (1999). Die Aufmerksamkeitsleistungen von 6-10-jährigen Kindern in der TAP. *Zeitschrift für Neuropsychologie*, 10, 87-102.
- Franzen, M. D., Smith, S. S., Paul, D. S. & MacInnes, W. D. (1993). Order effects in the administration of the Booklet Category Test and Wisconsin Card Sorting Test. *Archives of Clinical Neuropsychology*, 8, 105-110.
- Frey, A. (2006). *Validitätssteigerung durch adaptives Testen*. Frankfurt am Main: Peter Lang.
- Frey, A., Hartig, J. & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. *Diagnostica*, 55, 20-28.
- Hambros, K. (2002). On reasonableness of personality inventories with dichotomous item response format. *Psychologische Beiträge*, 44, 126-135.
- Hamilton, J. C. & Shuminsky, T. R. (1990). Self-awareness mediates the relationship between serial position and item reliability. *Journal of Personality and Social Psychology*, 59, 1301-1307.

- Hartig, J. & Hinrichs, J. (2002). *Kontexteffekte bei der Beantwortung von Persönlichkeitsfragebögen: Eine Überprüfung der Bedeutungsänderungshypothese*. Vortrag auf dem 43. Kongress der Deutschen Gesellschaft für Psychologie, Berlin.
- Hartig, J., Hölzel, B. & Moosbrugger, H. (2007). A Confirmatory Analysis of Item Reliability Trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research*, 42, 157-183.
- Heckhausen, J. & Heckhausen, H. (2010). *Motivation und Handeln* (4. Auflage). Berlin: Springer.
- Hergovich, A. (1994). *Feldabhängigkeit*. Frankfurt am Main: Lang.
- Hippler, H.-J. & Schwarz, N. (1987). Response effects in surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Hrsg.), *Social information processing and survey methodology* (S. 102-112). New York: Springer.
- Hofmann, K. & Kubinger, K. D. (2001). Herkömmliche Persönlichkeitsfragebogen und Objektive Persönlichkeitstests im „Wettstreit“ um Unverfälschbarkeit. *Report Psychologie*, 26, 298-304.
- Hogg, M. A. & Vaughan, G. M. (2008). *Social Psychology* (5. Aufl.). Harlow, UK: Pearson.
- Holocher-Ertl, S., Kubinger K. D. & Menghin S. (2005). *Big Five Plus One Persönlichkeitsinventar (B5PO)* [Software und Manual]. Mödling: Schuhfried.
- Hornke, L. F., Etzel, S. & Küppers, A. (2000). Konstruktion und Evaluation eines adaptiven Matrizentests. *Diagnostica*, 46, 182-188.
- Hornke, L. F., Etzel, S. & Rettig, K. (2010). *Adaptiver Matrizentest* [Software und Manual]. Mödling: Schuhfried
- Hossiep, R., Paschen, M. & Mühlhaus, O. (2003). *Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP)* (2. Aufl.). Göttingen: Hogrefe.

- James, W. H. (1957). *Internal versus external control of reinforcement as a basic variable in learning theory*. Unveröffentlichte Dissertation, Ohio State University, Columbus.
- Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.
- Kaplan, E. F., Goodglass, H. & Weintraub, S. (1983). *The Boston Naming Test* (2. Aufl.). Philadelphia: Lea & Febiger.
- Karner, T. (2002). The volunteer effect of answering personality questionnaires. *Psychologische Beiträge*, 44, 42-49.
- Kersting, M. (2003a). DIN 33430 – Entstehungsprozess, Ziele und Inhalte des neuen Qualitätsstandards für berufsbezogene Eignungsbeurteilung. *DGP Informationen*, 48, 2-6.
- Kersting, M. (2003b). Augenscheinvalidität. In K.D. Kubinger & R.S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 54-55). Weinheim: Beltz.
- Khorramdel, L. (2010). *Context effects in personnel selection: The influence of different contexts on the self-description in personality questionnaires as well as on test scores of cognitive ability tests and objective personality tests*. Unveröffentlichte Dissertation, Universität Wien.
- Khorramdel, L & Frebort, M. (2011). Context effects on test performance: What about test order? *European Journal of Psychological Assessment*, 27, 103-110
- Khorramdel, L. & Maurer, M. (unpubliziert). *WSP – Wiener Studieneignungs-Persönlichkeitsinventar*. Alle Rechte bei: Test- und Beratungsstelle, Institut für Entwicklungspsychologie und Psychologische Diagnostik, Universität Wien.
- Kieweg, D. (2004). *Testanordnung zur Überprüfung der Reflexions- und Modifikationsbereitschaft hinsichtlich des Selbstbildes*. Unveröffentlichte Diplomarbeit, Universität Wien.
- Kinsman, R. A. & P. C. Weiser (1976). Subjective symptomatology during work and fatigue. In E. Simonson & P. C. Weiser (Hrsg.), *Psychological aspects and physiological correlates of work and fatigue* (S. 336–405). Springfield, IL: Charles C. Thomas.

- Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, 55, 312–320.
- Knowles, E. S., Coker, M. C., Cook, D. A., Diercks, S. R., Irwin, M. E., Lundeen, E. J., Neville, J. W. & Sibicky, M. E. (1992). Order effects within personality measures. In N. Schwarz & S. Sudman (Hrsg.), *Context effects in Social and Psychological Research* (S. 221-236). New York: Springer.
- Kornmann, R. (2003). Förderungsorientierte Diagnostik. In K.D. Kubinger & R.S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologische Diagnostik* (S. 150-156). Weinheim: Beltz.
- Kraut, A. I., Wolfson, A. D. & Rothenberg, A. (1975). Some effects of position on opinion survey items. *Journal of Applied Psychology*, 60, 774-776.
- Krampen, G., Hense, H. & Schneider, J. F. (1992). Reliabilität und Validität von Fragebogenskalen bei Standardreihenfolge versus inhaltshomogener Blockbildung ihrer Items. *Zeitschrift für experimentelle und angewandte Psychologie*, 39, 229-248.
- Krampen, G. (1993). Effekte von Bewerbungsinstruktionen und Subskalenextraktion in der Fragebogendiagnostik. *Diagnostica*, 39, 97-108.
- Kubinger, K. D. (1993). Testtheoretische Probleme der Computerdiagnostik. *Zeitschrift für Arbeits- und Organisationspsychologie. Themenheft Computerdiagnostik*, 11, 130-137.
- Kubinger, K. D. (2002). On faking personality inventories. *Psychologische Beiträge*, 44, 10-16.
- Kubinger, K. D. & Jäger, R.S. (Hrsg.) (2003). *Schlüsselbegriffe der Psychologischen Diagnostik*. Weinheim: Beltz/PVU.
- Kubinger, K. D. (2009). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens* (2. Aufl.). Göttingen: Hogrefe.
- Kubinger, K. D. & Ebenhö, J. (2007). *Arbeitshaltungen – Kurze Testbatterie: Anspruchsniveau, Frustrationstoleranz, Leistungsmotivation, Impulsivität/Reflexivität* [Software und Manual]. Mödling: Schuhfried.

- Kubinger, K. D. & Hofmann, K. (1998). Die Objektive Persönlichkeitstestbatterie „Arbeitshaltungen“ aus der Sicht der „Big Five“. In J. Glueck, O. Vitouch, M. Jirasko & B. Rollett (Hrsg.), *Perspektiven Psychologischer Forschung in Österreich - Band 2*. Wien: WUV Universitätsverlag, 189 – 192.
- Kubinger, K. D. & Litzenberger, M. (2003). Zur Validität der Objektiven Persönlichkeits-Test-Batterie „Arbeitshaltungen“. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 24, 119-133.
- Kubinger, K. D., Rasch, D. & Yanagida, T. (2011). *Statistik in der Psychologie – vom Einführungskurs bis zur Dissertation*. Göttingen: Hogrefe.
- La Rue, A. & Olejnik, A. B. (1980). Cognitive “priming” of principled moral thought. *Personality and Social Psychology Bulletin*, 6, 413-416.
- Leary, L. F. & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, 55, 387-413.
- Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: PVU.
- Liepmann, D., Beauducel, B., Brocke, B. & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000-R (IST 2000-R)* (2. Aufl.). Göttingen: Hogrefe
- Linden, W. J. van der & Glas, C. A. W. (Hrsg.) (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht, Boston, London: Kluwer Academic Publishers.
- Llorente, A. M., Sines, M. C., Rozelle, J. C., Turcich, M. R. & Casatta, A. (2000). Effects of test administration order on children’s neuropsychological performance: Emerging one-word expressive and receptive language skills. *The Clinical Neuropsychologist*, 14, 162–172.
- Makransky, G. & Glas, C. A. W. (2013). The Applicability of Multidimensional Computerized Adaptive Testing for Cognitive Ability Measurement in Organizational Assessment. *International Journal of Testing*, 13, 123-139.
- Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, 5, 105-121.

- Menghin, S. & Kubinger, K. D. (1996). Zur Legende: „Testpersonen beantworten dem Computer persönliche und intime Fragen offener als einem Testleiter“ – Ergebnisse eines Experiments. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 17, 163-169.
- Moosbrugger, H. & Kelava, A. (2007). *Testtheorie und Fragebogenkonstruktion* (2. Aufl.). Heidelberg: Springer.
- Moosbrugger, H. & Oehlschlägel, J. (1996). *Frankfurter Aufmerksamkeits-Inventar (FAIR)*. Bern: Huber.
- Möseneder, D. & Ebenhöf, J. (1996). *Ein Simulationstest zur Erfassung des Entscheidungsverhaltens (ILICA)* [Software und Manual]. Frankfurt: Swets.
- Muller, K. E. & Peterson, B. L. (1984). Practical methods for computing power in testing the multivariate general linear hypothesis. *Computational Statistics and Data Analysis*, 2, 143-158.
- Muscio, B. (1921). Is a fatigue test possible? *British Journal of Psychology*, 12, 31–46.
- Myers, C. S. (1937). Conceptions of mental fatigue. *American Journal of Psychology*, 50, 296–306.
- Neuger, G. J., O’Leary, D. S., Fishburne, F. J., Barth, J. T., Berent, S., Giordani, B. & Boll, T. J. (1981). Order effects on the Halstead-Reitan Neuropsychological Test Battery and allied procedures. *Journal of Consulting and Clinical Psychology*, 49, 722-730.
- Ostendorf, F. & Angleitner A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae (NEO-PI-R)*. Göttingen: Hogrefe.
- Oswald, W. D. & Rothe, E. (1987). *Der Zahlen-Verbindungs-Test (ZVT)*. Göttingen: Hogrefe.
- Ortner, T. M. (2001). *Ein psychophysiologischer Validierungsansatz zur Belastbarkeitstestbatterie BAcO*. Unveröffentlichte Diplomarbeit, Universität Wien.
- Ortner, T. M. (2004). On changing the position of items in personality questionnaires: Analyzing effects of item sequence using IRT. *Psychology Science*, 46, 466–476.

- Ortner, T. M. (2008). Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment*, 16, 249-257.
- Ortner, T. M., Horn, M., Kersting, M., Kubinger, K. D., Schmidt-Atzert, L., Proyer, R. T., Schuhfried, G., Schütz, A., Wagner-Menghin, M. & Westhoff, K. (2007). Standortbestimmung und Zukunft Objektiver Persönlichkeitstests. *Report Psychologie*, 32, 60-69.
- Ortner, T. M., Kubinger, K. D., Schrott, A., Radinger, R. & Litzenberger, M. (2006). *Belastbarkeits-Assessment: computerisierte Objektive Persönlichkeits-Testbatterie – Deutsch (BAcO-D)* [Software und Manual]. Frankfurt am Main: Harcourt Test Services.
- Ortner, T. M., Kubinger, K. D., Schrott, A., Radinger, R. & Litzenberger, M. (2011). *Belastbarkeits-Assessment BACO* [Software und Manual]. Mödling: Schuhfried.
- Ortner, T. M., Proyer, R. T. & Kubinger, K. D. (Hrsg.) (2006). *Theorie und Praxis Objektiver Persönlichkeitstests*. Bern: Huber
- Perlini, A. H., Lind, D. L. & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology/Psychologie canadienne*, 39, 299-307.
- Petty, R. E. & Cacioppo J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.
- Proyer, R. T. & Ortner, T. M. (2004). Zum Einfluss der Einschätzung des Testmaterials durch Testpersonen bei der Arbeit mit Objektiven Persönlichkeitstests. Unveröffentlichter Forschungsbericht aus den Mitteln des Theodor-Körner-Fonds: Wien.
- Rasch, D. & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science*, 46, 175-208.
- Raven, J., Raven, J. C. & Court, J. H. (2008). *Standard Progressive Matrices (SPM)*. Mödling: Schuhfried.

- Reimann, G. (2010). Rechtliche Rahmenbedingungen. In K. Westhoff, L.J. Hellfrisch, L.F. Hornke, K.D. Kubinger, F. Lang, H. Moosbrugger, A. Püschel & G. Reimann (Hrsg.), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430*. Lengerich: Pabst.
- Rollett, B. & Bartram, M. (1998). *Anstrengungsvermeidungstest (AVT)* (3. Aufl.). Göttingen: Hogrefe.
- Rost, D. H. & Hoberg, K. (1997). Itempositionsveränderungen in Persönlichkeitsfragebögen: Methodischer Kunstfehler oder tolerierbare Praxis? *Diagnostica*, 43, 97-112.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Verlag Hans Huber.
- Ryan, J. J., Glass, L. A., Hinds, R. M. & Brown, C. N. (2010). Administration order effects on the Test of Memory Malingering. *Applied Neuropsychology*, 17, 246–250.
- Schaarschmidt, U. (2003). Diagnostik im Personalwesen. In K.D. Kubinger & R.S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 317-324). Weinheim: Beltz.
- Schriesheim, C. A. (1981). Leniency effects on convergent and discriminant validity for grouped questionnaire items: A further investigation. *Educational and Psychological Measurement*, 41, 1093–1099.
- Schriesheim, C. A., Kopelman, R. E. & Solomon, E. (1989). The effect of grouped versus randomized questionnaire format on scale reliability and validity: A three-study investigation. *Educational and Psychological Measurement*, 49, 487–508.
- Schmidtke, H. (1976). Disturbance of processing of information. In E. Simonson & P. C. Weiser (Hrsg.), *Psychological aspects and physiological correlates of work and fatigue* (S. 336–405). Springfield, IL: Charles C Thomas.
- Schneider, H. & Stein, D. (2006). Personalpolitische Strategien deutscher Unternehmen zur Bewältigung demografisch bedingter Rekrutierungsengpässe bei Führungskräften. *IZA Research Report*, 6. Bonn: Institut zur Zukunft der Arbeit.

- Schneider, M. & Stern, E. (2010). The developmental relations between conceptual and procedural knowledge: A multimethod approach. *Developmental Psychology*, 46, 178-192.
- Schuler, H. (1996). *Psychologische Personalauswahl: Einführung in die Berufseignungsdiagnostik*. Göttingen: Verlag für Angewandte Psychologie.
- Schuler, H. (2006). Arbeits- und Anforderungsanalyse. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 45-68). Göttingen: Hogrefe.
- Schuler, H. & Höft, S. (2006). Konstruktorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 102-144). Göttingen: Hogrefe.
- Schwarz, N., Hippler, H.-J., & Noelle-Neumann, E. (1992). Cognitive model of response order effects. In N. Schwarz & S. Sudman (Hrsg.), *Context effects in Social and Psychological Research* (S. 187-201). New York: Springer.
- Schwarz, N. & Sudman, S. (Hrsg.) (1992). *Context effects in Social and Psychological Research*. New York: Springer.
- Seo, T., Kanda, T. & Fujikoshi, Y. (1995). The effects of nonnormality on tests for dimensionality in canonical correlation and MANOVA models. *Journal of Multivariate Analysis*, 52, 325-337.
- Smith, T. W. (1992). Thoughts on the nature of context effects. In N. Schwarz & S. Sudman (Hrsg.), *Context effects in Social and Psychological Research* (S. 163-184). New York: Springer.
- Sparfeldt, J. R., Schilling, S. R., Rost, D. H. & Thiel, A. (2006). Blocked versus randomized format of questionnaires – A confirmatory multigroup analysis. *Educational and Psychological Measurement*, 66, 961–974.
- Strack, F. (1992). Order effects in survey research: Activation and information functions of preceding questions. In N. Schwarz & S. Sudman. (Hrsg.), *Context effects in Social and Psychological Research* (S. 23-34). New York: Springer.
- Strack, F. (1994). Zur Psychologie der standardisierten Befragung: Kognitive und kommunikative Prozesse. In D. Albert, K. Pawlik, K.H. Stapf & W. Stroebe (Hrsg.), *Lehr- und Forschungstexte Psychologie*, 48. Berlin: Springer Verlag.

- Tabachnick, G. & Fidell, S. (2007). *Using Multivariate Statistics* (5. Aufl.) Boston, MA: Pearson.
- Testkuratorium (1986). Beschreibung der einzelnen Kriterien für die Testbeurteilung. *Diagnostica*, 32, 358-360.
- Thissen, D. & Mislevy, R. J. (2000). Testing Algorithms. In H. Wainer (Hrsg.), *Computerized Adaptive Testing: A primer* (S. 101-133) New Jersey: Lawrence Erlbaum Associates.
- Thorndike, E. L. (1900). Mental fatigue. *Psychology Review*, 7, 547-579.
- Tombaugh, T. N. (1996). *Test of Memory Malingering (TOMM)*. New York: Multi Health Systems.
- Tonidandel, S., Quiñones, M. A. & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, 87, 320–332.
- Tourangeau, R. (1984). Cognitive science and survey methods. In T. Jabine, M. Straf, J. Tanur & R. Tourangeau (Hrsg.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (S. 73-100). Washington, DC: National Academy Press.
- Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103, 299-314.
- Tulsky, D. S. & Zhu, J. (2000). Could test length or order affect scores on Letter Number Sequencing of the WAIS-III and WMS-III? Ruling out effects of fatigue. *The Clinical Neuropsychologist*, 14, 474-478.
- Wagner-Menghin, M. (2003). Computerdiagnostik. In K.D. Kubinger & R.S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik*. Weinheim: Beltz.
- Wainer, H. & Eignor, D. (2000). Caveats, Pitfalls, and Unexpected Consequences of implementing Large Scale Computerized Testing. In Wainer, H. (Hrsg.), *Computerized Adaptive Testing: A primer* (S. 271-299). Hillsdale, NJ: Lawrence Erlbaum Associates.

- Westhoff, K., Hellfritsch, L. J., Hornke, L. F., Kubinger, K. D., Lang, F., Moosbrugger, H., Püschel, A. & Reimann, G. (Hrsg.) (2010). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430*. Lengerich: Pabst.
- Wild, B. (1986). *Der Einsatz adaptiver Teststrategien in der Fähigkeitsmessung*. Unveröffentlichte Dissertation, Universität Wien.
- Woike, J. K. (2003). Selektionsorientierte Diagnostik. In K.D. Kubinger & R.S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologische Diagnostik* (S. 377-380). Weinheim: Beltz.
- Wottawa, H. & Hossiep, R. (1997). *Anwendungsfelder psychologischer Diagnostik*. Göttingen: Hogrefe.
- Wottawa, H. (2002). Einige wichtige Entwicklungen der Psychologischen Diagnostik im letzten Jahrzehnt. *Psychologie in Österreich*, 22, 3-5.
- Wechsler, D. (1997). *Wechsler Memory Scale – Third edition*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1997). *Manual for the Wechsler Adult Intelligence Scale - Third edition*. San Antonio, TX: The Psychological Corporation.
- Zimmermann, P. & Fimm, B. (1993). *Testbatterie zur Aufmerksamkeitsprüfung (TAP)*. Freiburg: Psytest.
- Zhu, J. & Tulskey, D. S. (2000). Co-norming the WAIS-III and WMS-III: Is there a testorder effect on IQ and memory scores? *The Clinical Neuropsychologist*, 14, 461-467.

8. Lebenslauf

Alexander Leonard Schünemann

ZUR PERSON

Geburtsdatum: 18-06-1987
Geburtsort: Hamburg
Nationalität: Deutsch
Email: leonard.schuenemann@univie.ac.at

AUSBILDUNG

Seit 2007 Diplomstudium der Psychologie an der Universität Wien

1993 – 2006 Abitur am Kepler Gymnasium – Freiburg im Breisgau
Jürgen-Fuhlendorf-Gymnasium und Grundschule – Hamburg

BERUFSERFAHRUNG

2013 Forschungsaufenthalt – Projekt zur biologischen Bewegungswahrnehmung
University of Aberdeen, Schottland

2012 – 2014 Studienassistent am Arbeitsbereich Psychologische Diagnostik
Universität Wien

2012 – 2013 Projektarbeit Testentwicklung – IMPULS|2 Professional
Humanware GmbH, Wien

2011 Praktikum an der Test- und Beratungsstelle des Arbeitsbereichs Psychologische Diagnostik
Universität Wien

2010 – 2012 Mitarbeiter der Thalia Buch und Medien GmbH, Wien

2008 – 2010 Mitarbeiter der Raiffeisen Banken und Daten Service Center GmbH, Wien

DIVERSES

Konferenzen ECPA, Riga (2011) – ICT, Amsterdam (2012) – TEAP, Wien und DPPD, Greifswald (2013)

Vorträge Schünemann, L. & Khorramdel, L. (2013). Testreihenfolge-Effekte bei Persönlichkeits- und Leistungsdiagnostik in realen Auswahl-situationen: drei Experimente in der Personalauswahl. *12. Arbeitstagung der Fachgruppe Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik*, Greifswald, 23.-25. September 2013.

 Schünemann, L. & Khorramdel, L. (2013). Test order effects within a computer test battery in high stakes assessment: varying objective personality tests and cognitive ability tests. *55. Tagung experimentell arbeitender Psychologen*, Wien, 24.-27. März 2013.

Fremdsprachen Englisch, Französische, Plattdeutsch (genetisch bedingte erweiterte Grundkenntnisse)

